DOCUMENT RESUME

ED 139 050 CS 501 696

AUTHOR Krippendorff, Klaus
TITLE Clustering.
PUB DATE Apr 77
NOTE 64p.; Paper presented at the Multivariate Analysis in Communication Research Conference (Pacific Grove, California, April 3-6, 1977) 1977)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Classification; *Cluster Analysis; Cluster Grouping; *Communication (Thought Transfer); *Educational Research; *Grouping Procedures; *Statistical Analysis; Validity

ABSTRACT

Clustering techniques seek to group together objects or variables that share some observed qualities or, alternatively, to partition a set of objects or variables into mutually exclusive classes whose boundaries reflect differences in the observed qualities of their members. This paper reviews the general principles underlying clustering techniques and discusses the various options facing the analyst who wishes to make use of them. Topics considered include the canonical form of the data, relevance of the variables included in the analyses, the choice of bi-ordinal or multi-ordinal clustering techniques, the use of distance authoritarianism were related but discriminable entities. Authoritarianism was not analysis, alternative ways of displaying results, measures of relationships between clusters, clustering algorithms, and procedures for validating clusters which emerge from the statistical procedures. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED139050

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CLUSTERING

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Klaus Krippendorff

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

by

Klaus Krippendorff

The Annenberg School of Communications
University of Pennsylvania, Philadelphia

October, 1976
revised May, 1977

Prepared for Multivariate Techniques of Analysis in Communication Research, Cappella, Joseph N., and Monge, Peter R.(Eds.), New York: Academic Press, 1978

Contents

Clustering	1
Canonical Form of Data	3
Relevance of Data	4
Ordinality of Data	6
Derived Form of Data: Distances	9
Goals and Computational Efforts	17
Presentation of Results	20
Properties of Emerging Clusters	27
Clustering Algorithms	37
Validation and Vindication	51
References	58

CS501696

Clustering

Clustering seeks to group or to lump together objects or variables that share some observed qualities or, alternatively to partition or to divide a set of objects or variables into mutually exclusive classes whose boundaries reflect differences in the observed qualities of their members. Thusly typologies emerge from data which in turn represent a reduction of data complexity and lead to conceptual simplifications.

Clustering should not be confused either with the analysis of the groupings made by a data source or with the problems of assigning objects to the categories in which they belong. Clusters emerge from the interaction between the characteristics that are manifest in multivariate data and the assumptions that are built into the procedure. The recognition of these assumptions pertains to problems of validity, to which much of the paper and its conclusion is devoted.

Clustering originated in anthropology (Driver and Kroeber, 1932) and in psychology (Zubin, 1938; Tryon, 1939) as a response to the then common scientific aim to develop typologies of cultures and of individuals in terms of the characteristics they exhibit. Computational problems inhibited the development of these ideas until the early 1960's where interest into such techniques emerged in a variety of disciplines, notably in biology (Sokal and Sneath, 1963). Applications are now so numerous that references to them would fill a book (see Sneath and Sokal, 1973). Problems to which clustering has found answers range from counting dust particles and bacteria, to land allocation in urban planning and political campaigning. Among the disciplines in which clustering has been successful is psychology, anthropology, sociology, political science, economics, management, geography, and literature, - virtually the whole spectrum of the behavioral and social sciences (Bailey, 1974) in which data do not exhibit the determinism of the natural sciences and theories are based on types, categories and differentiations that knowingly omit some of the less significant variations.

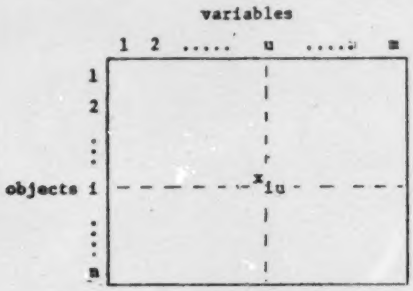
In communication research, clustering provides a valuable tool for identifying

cliques from sociometric or communication network type data, for example, or for detecting "invisible colleges" as manifest through citations of literature in scientific publications. Clustering is also used for grouping concepts that appear highly associated in given messages into stereotypes, for developing emic type categories for content analysis from receiver responses or for detecting redundant questionnaire items that may be explained by a common underlying variable. Clustering may also be used to simplify the representation of complex communication systems and thus provides the pretext for other forms of analysis including modelling.

Just as the application has spread over many disciplines so have available clustering procedures grown in variety. In the following I will not attempt to present a survey of either. Rather, based on the belief that all clustering techniques follow a few basic principles, with ample room open for further applications and developments of details, I will discuss some of the options an analyst faces when deciding among existing clustering procedures or when assembling one for his special purpose and I will discuss some of the implications such choices have regarding computational efforts, validity and interpretability of results. The paper provides in a sense a collection of tools for evaluating what exists and for constructing anew what is needed when multivariate data are to be analyzed by what has become known as clustering.

Canonical Form of Data

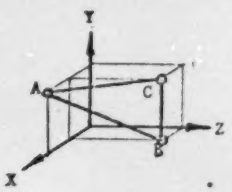
The canonical form of data for clustering is an n×m matrix X with entries x_{iu} denoting measures of some sort. The common interpretation of this matrix is that each of n objects is described in terms of m values, each pertaining to a different variable. The rows of this matrix are m-tuples or vectors with m components:

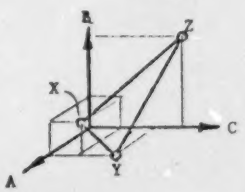


The variables may have different metrics: nominal, ordinal, interval and ratio metric, and may have any degree of freedom. This includes binary variables as a special case.

It is basic to clustering that a matrix X whose variables have the same metric throughout can be interpreted in two ways, as an Objects×Variables-matrix and as a Variables×Objects-matrix for it is then possible to cluster objects in terms of variables, variables in terms of objects and indeed both in terms of each other. For example, the following matrix can be depicted in either way:

		variables X	Y	Z
objects	A	2	2	0
	B	1	0	3
	C	1	2	3





In the left diagram objects B and C are relatively close while in the right diagram variables X and Y are relatively close.

An important distinction is whether data are complete or not. If some of the entries x_{iu} are not known or unavailable for analysis, data are incomplete and require that special assumptions be adopted to compensate for the missing entries. This paper considers complete data only.

Relevance of Data

It is fundamental that the variables that are chosen as descriptive of the objects in a sample be relevant to the attempted clustering. Individuals can be categorized in terms of their income, occupation and social status, in terms of psychopathologies, in terms of physical conditions, their weight, height, strength, in terms of their life styles, etc. The choice depends on the purpose of the clusters that is expected to emerge. While statistical criteria for choosing among the objects of a population are available in the form of sampling theory, criteria for choosing among the potentially infinite universe of possible variables are not so clear.

Generally, variables that are constant in the sample and those whose occurrence is random may be ignored in cluster analysis as they contribute nothing to or merely pollute differentiations among objects respectively. Also redundant variables (different measures of the same underlying dimension) should be avoided for they only increase computational efforts. The task of identifying constant, random and redundant variables is a purely syntactical one and can be accomplished by a variety of analytical techniques. For example, factor analysis has been used to identify orthogonal variables that are in a sense least redundant.

Beyond the elimination of constant, random and redundant variables the researcher requires a theory or at least good intuition to decide on the relevance of variables to a given problem. For example, if it were to be the task to cluster psychotherapeutic patients for simplifying their clinical treatment, the researcher will want to derive from existing theories all variables pertaining to such treatment. Or if it were to be the task to develop typologies for social organization of industrial enterprises the researcher will want to be sure that none of the variables that sensible writers have associated with industrial organizations are excluded. One of the typical problems of deriving measurement variables from various theoretical writings is that variables may pertain to different levels of aggregation, for example, alienation, an individual's state, vs. vertical organization, a characteristic of formal

structure in which individuals take part. Another is that theoretical concepts are often abstract and multidimensional. For example authoritarianism may not be measurable by a single variable. Finally, variables may not have the same weight. For example, while "education" is more important than "sex" in explaining the formation of social groups focussing on academic topics, the situation is reversed when it comes to humor, social stereotypes and economic expectations. Ideally, all variables chosen are on the same level of analysis, equal in weight, logically independent of each other and feed into a theory or conceptual system that renders the description of the objects of analysis meaningful specimens.

The lack of relevance is an important source of the difficulty of interpreting clustering results. But, since relevance is extraneous to the clustering procedure it can only be mentioned here. The other source of difficulty is the validity of a clustering procedure to which another section of this paper is devoted.

Ordinality of Data

Clustering, just as any multivariate technique, is responsive to and ultimately accounts for certain patterns that are manifest in given data. Just how these patterns are defined, recognized and represented distinguishes between kinds of multivariate techniques.

A common justification for grouping several objects into one cluster is their relative proximity. Distances, differences, proximities, similarities, disagreements, correlations, etc. are quantitative expressions for relations between two objects or between two variables. Distances, etc., having exactly two arguments, belong to the class of binary relations. Multivariate techniques that are based on binary relations are bi-ordinal (of order two). Bi-ordinal clustering procedures yield clusters all of whose objects stand to each other in a certain pairwise relationship, proximity being one such relationship.

That there exist more complex relations that cannot be decomposed without loss into a set of binary relations is easily demonstrated and provides the ground for the distinction between bi-ordinal and multi-ordinal clustering techniques.

Suppose one is given a 3 by 4 matrix X with binary attributes, 0 and 1, in all four variables:

variables

	1	2	3	4
A	1	1	0	0
B	1	0	1	0
C	0	1	1	0

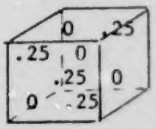
The three 2-dimensional contingency matrices of probabilities within which any distance measure or association between pairs of objects is to be defined are:

A \ B	0	1
0	.25	.25
1	.25	.25

A \ C	0	1
0	.25	.25
1	.25	.25

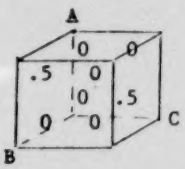
B \ C	0	1
0	.25	.25
1	.25	.25

They would indicate that the matching of attributes is entirely due to chance. A bi-ordinal clustering technique would therefore find no justification for merging any two objects into a cluster. However, when one considers the values associated with the three objects, not in pairs, but as a triple, one finds a strong tertiary relation present. The three-dimensional contingency matrix reveals this graphically:



If a higher-order relation holds among three or more objects, as in this example, this might well be an important reason for clustering these objects despite the absence of binary relations.

An example of a relation between three objects that is fully explainable in terms of any two of its three component binary relations is this:



	B	
A	.5	0
	0	.5

	C	
A	.5	0
	0	.5

	C	
B	.5	0
	0	.5

Here, bi-ordinal clustering would be perfectly justifiable for their is nothing unique about the combination of the three objects that could not be described in binary terms.

The point of this demonstration is that relations of lower ordinality do not imply anything about the presence or absence of relations of higher ordinality. Data in the form of distances or similarities can feed bi-ordinal techniques only. Multi-ordinal clustering techniques require continuous interaction with data in their canonical form.

The significance of higher order relations in reality and inherent in multivariate data cannot be underestimated. For example, when social groups are formed from individuals it is well known that the behavior of a larger group cannot be inferred from the knowledge of the behavior within pairs of individuals. There often are great quali-

tative differences between two-person interaction and threeperson interaction, coalition formation being just one example requiring three or more parties. Even in chemistry, most man-made substances emerge when a certain number and quantity of elementary substances meet under suitable conditions. What does emerge cannot be predicted from known effects that any pair of elementary substances may have on each other. Or, if the components of an electronic equipment have more than one input and/or more than one output, their behavior can no longer be described by a two valued function. The switching network is then more complex precisely because the whole is different from the sum of the input-output relations of its parts. The difference between a whole and its parts is variously called organization, synergy, interaction effects, or a Gestalt and points directly to the difference between binary and higher-order relations. Unfortunately not all multivariate techniques are capable of exploring the multi-ordinal character of the data to which they are applied.

The choice between bi-ordinal and multi-ordinal clustering techniques should be made dependent on the nature of the data. Whenever objects form natural clusters on the basis of wholistics qualities similar to those mentioned above, multi-ordinal techniques are called for. If the choice of a bi-ordinal technique is dictated by their availability, the researcher should at least be aware of or measure in quantitative terms what his clusters will omit.

Derived Form of Data: Distances

Bi-ordinal clustering usually starts with data in form of a square matrix D whose entries d_{ij} measure some distance, difference or dissimilarity either (and generally) between all pairs of objects or (provided they possess the same metric) between all pairs of variables:

objects

	1	2	....	j	...	n
1	d_{11}	d_{12}	- - - -	d_{1j}	- - - -	d_{1n}
i	d_{i1}	d_{i2}	- - - -	d_{ij}	- - - -	d_{in}
:	:	:		:		:
:	:	:		:		:
n	d_{n1}	d_{n2}	- - - -	d_{nj}	- - - -	d_{nn}

Distances between the same objects must be zero, and positive and symmetrical otherwise:

$$d_{ii} = 0$$
$$d_{ij} \geq d_{ii}$$
$$d_{ij} = d_{ji}$$

In order to possess at least interval metric properties within many-dimensional space, such derived forms of data must also satisfy the triangle inequality:

$$d_{ik} \leq d_{ij} + d_{jk}$$

and in some cases the ultrametric inequality (Jardine and Gibson, 1971; Johnson, 1967):

$$d_{ik} \leq \max(d_{ij}, d_{jk})$$

Data may also be represented in terms of similarities, agreements, measures of resemblance or correlation, s_{ij}. Similarities and distances are inversely related with the least similar objects giving rise to large distances and small distances reflecting strong resemblances. Similarity measures are usually convertable into distance measures, for example by:

$$d_{ij} = s_{ii} - s_{min} - s_{ij}$$

$$d_{ij} = \sqrt{s_{ii} - s_{min} - s_{ij}}$$

$$d_{ij} = \frac{s_{ii} - s_{ij}}{s_{ii} - s_{min}}$$

$$d_{ij} = \sqrt{(s_{ii} - s_{min})(s_{ii} - s_{ij})}$$

The latter is a generalization from Gower (1966) who developed this form to convert similarity measures that range between -1 and +1. A useful conversion for correlation coefficients has been suggested by Tukey:

$$d_{ij} = 1 - r_{ij}^2$$

It expresses the degree to which two objects are linearly related (positive or negative) and takes its maximum value when the objects are statistically independent. This serves as an example that distances and similarities can have many different interpretations which need to be understood before a clustering is attempted. Motivations for these conversion formula cannot be given here. The section on Relevance of Data applies here in part.

In the following I will give several distance and similarity measures and show how some of the more familiar measures can be regarded as special cases. For this purpose I will first define four kinds of differences, one for each metric, then present methods of standardizing such differences across different metrics and finally present a few key distance and similarity measures.

That difference notions depend on the metric of the variables involved may be seen in the comparison between nominal and interval scale properties. In nominal data, in which values mark qualitative distinctions without any implied order, one value matches with another or it does not and all differences between mismatching pairs of values must therefore be the same. In interval data, in which values represent quantities rather than qualities, differences can become a function of their algebraic difference and may be large or small. These intuitive notions can be given rigorous forms:

For unordered qualities or nominal scales and binary attributes, which is the special case of all other metrics, differences are manifest in mismatches and carry the same weight. The difference between two values x_{iu} and x_{ju} of the u-th variable is then:

$$\Delta_{ij,u} = \begin{cases} 0 & \text{iff } x_{iu} = x_{ju} \\ 1 & \text{iff } x_{iu} \neq x_{ju} \end{cases}$$

For ordinal variables in which merely the rank orders count, such a difference is a function of the number of ranks above and below the two values to be compared. With:

$$x^*_{iu} = 1 + \frac{1}{n}\left[\sum_{x_{ku} > x_{iu}} n_{x_{ku}} - \sum_{x_{ku} < x_{iu}} n_{x_{ku}}\right]$$

and x^*_{ju} defined analoguously, the difference in ordinal scales becomes:

$$\Delta_{ij,u} = |x^*_{iu} - x^*_{ju}|$$

For variables with interval metric the difference is as discussed above:

$$\Delta_{ij,u} = |x_{iu} - x_{ju}|$$

and for variables with ratio metric, the difference may be expressed by:

$$\Delta_{ij,u} = \frac{|x_{iu} - x_{ju}|}{|x_{iu}| + |x_{ju}|}$$

If both x_{iu} and x_{ju} are positive, as should be expected in ratio-level measurements, then the ratio difference is a modification of Lance and Williams' (1967) Canberra-metric, All of these differences are taken from Krippendorff (1973) where also the motivations for their form are found.

into a measure of distance

The most obvious form of agregating differences across variables is by summing some power r of it:

$$d_{ij} = \sqrt[r]{\sum_{u=1}^{m} (\Delta_{ij,u})^r}$$

for r=2, d_{ij} corresponds to the familiar Euclidian distance in multi-dimensional space which has been used since Heincke (1898), is common in research on the semantic differential, (Osgood, Suci and Tannenbaum, 1957) and has been discussed in the cluster analysis literature by Sokal and Sneath (1963), Gower and Ross (1969) and many others. While r=1 assigns equal weights to differences $\Delta_{ij,u}$, an increase in the exponent r increases the impact of larger differences over smaller ones, and thereby affects the nature of the clusters formed.

The Euclidian distance is appropriate when all variables possess the same metric but not when the metric of the variables differs. Variables then will have to be standardized. There seem to be three methods of standardizing distances across different metrics.

The first is a reduction of the power of the metric of all variables to the least powerful metric among the variables. The possible metrics listed in the order of least to most powerful is:

nominal
ordinal
interval
ratio

Thus if there are ratio scales (e.g. numerical age, income in $) and ordinal scales (e.g. using such values as most conservative, somewhat conservative, neutral,...) the values of all variables would then have to be regarded merely by their rank within the set of all values. Or when binary attributes occur, all variables would then have to be dichotomised (e.g. above or below a certain age, income, liberal vs. conservative).

A second method is to transform all measures x_{iu} so that their range falls within the interval 0 and 1. With $\min_u(x_{ku})$ and $\max_u(x_{ku})$, k=1,...,n, the smallest and largest value in the u-th variable respectively, the transformed value is:

$$x'_{iu} = \frac{x_{iu} - \min_u(x_{ku})}{\max_u(x_{ku}) - \min_u(x_{ku})}, \quad k = 1, \ldots, n .$$

However, this method, which is proposed by Sokal and Sneath (1963) and Gower (1971), is inapplicable to nominal metric variables, and if applied to ordinal data, it would assume interval characteristics that are not there.

A third method, and one that I prefer, is to standardize the variance in each variable before summing. For this purpose one computes a variable-specific weight:

$$w_{u(r)} = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n}(\Delta_{ij,u})^r}$$

so that:

$$\partial_u^r = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{u(r)} (\Delta_{ij,u})^r = 1$$

for each variable. The weighted distance then follows the form proposed by Bock (1974):

$$d_{ij} = \sqrt[r]{\sum_{u=1}^{m} w_{u(r)} (\Delta_{ij,u})^r}$$

Evidently, when r=2 and all $w_{u(2)}=1$, the distance becomes again the Euclidian distance.

Another generalization of the Euclidian distance has been proposed by Mahalanobis (1936). For n>m, let the m*m covariance matrix Σ have entries:

$$\sigma_{uv} = \frac{1}{n} \sum_{i=1}^{n} (x_{iu} - \bar{x}_{.u})(x_{iv} - \bar{x}_{.v})$$

where $\bar{x}_{.u}$ and $\bar{x}_{.v}$ are the means in variable u and v respectively. Let the inverse of this matrix Σ^{-1} have entries σ^{uv}, then the Mahalanobis distance is defined by:

$$d_{ij} = \sqrt{\sum_{u=1}^{m} \sum_{v=1}^{m} \sigma^{uv} \Delta_{ij,u} \Delta_{ij,v}}$$

An important feature of the Mahalanobis distance is that it depends on all n objects simultaneously with its range in fact a function of n:

$$0 \leq d_{ij} \leq \sqrt{2n} \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2mn^2$$

The Mahalanobis distance also eliminates the effects of possible correlations among pairs of variables on the distance between two objects. And it is relatively independent of the ranges in each variable. The Mahalanobis distance is not applicable across different metrics, however. But this drawback might be correctable by standardizing the variance in each variable according to the third method discussed above. Accordingly, the corrected covariance matrix Σ^* has then entries:

$$\sigma^*_{uv} = w_{u(1)} w_{v(1)} \sigma_{uv}$$

and the corrected distance becomes:

$$d_{ij} = \sqrt{\sum_{u=1}^{m} \sum_{v=1}^{m} \sigma^{*uv} w_{u(1)} w_{v(1)} \Delta_{ij,u} \Delta_{ij,v}}$$

A most common similarity measure is the product-moment correlation coefficient r_{ij} which is defined by:

$$r_{ij} = \frac{\sum_u x_{iu}x_{ju} - \bar{x}_{i.}\bar{x}_{j.}}{(\sum_u x_{iu}^2 - \bar{x}_{i.}^2)(\sum_u x_{ju}^2 - \bar{x}_{j.}^2)}$$

Where $\bar{x}_{i.}$ and $\bar{x}_{j.}$ are the mean values for objects i and j respectively. The reliance on mean values across all variables makes an at-least-interval metric for all variables a requirement for its application.

While the correlation coefficient can easily be converted into a distance measure by one of the conversion formulae, its interpretation as a similarity measure is not too clear, however. $r_{ij}=0$ denotes statistical independence and $|r_{ij}|=1$ that a linear relation is present. Non-linear relationships, reduce the value of $|r_{ij}|$ and correlations between variables may increase its value. Whether the existence of a positive or negative linear relationship is sufficient to consider two objects similar must be decided in conjunction with the desired properties of the clusters to be computed. Opinions on the use of correlation coefficients for clustering vary considerably in the literature.

In an approach to multi-ordinal clustering of nominal data, Krippendorff (1969, 1974) used information theoretical measures to assess the loss of structure in many-dimensional spaces caused by the grouping of the objects' qualitative descriptions. Information theory is applicable to unordered values only. Any ordering within variables has no effect on the notion of structure. The notion of structure here considered may be paraphrased by "related differentiation," "trans-information," "multiple-order interaction," "relational entropy" and might be said to be the opposite of redundancy and randomness. The loss in the amount of structure due to the elimination of qualitative distinctions in one or more variables can be expressed by several distances between two objects of which I will define here only one. Let $n_{<i>}$ be the number of objects with the description $<x_{i1},\ldots,x_{iu},\ldots,x_{im}>$, $n_{x_{.u}}$ be the number of objects described in terms of x in the u-th variable, and $n_{x_{iu}}$ be the number of objects that share the value x in the u-th variable with the object i. The total amount of structure within

the m-dimensional space, the total amount of relatedness manifest in the m-valued distribution of objects is:

$$T = \sum_{i} \frac{n_{<i>}}{n} \log_2 \frac{n_{<i>}}{n} - \sum_{u=1}^{m} \sum_{x} \frac{n_{x.u}}{n} \log_2 \frac{n_{x.u}}{n}$$

And the distance between two objects i and j becomes the loss in structure when all hyperplanes within which i and j is located were to be merged. To express this, the two m-tuples $<x_{i1},\ldots,x_{iu},\ldots,x_{im}>$ and $<x_{j1},\ldots,x_{ju},\ldots,x_{jm}>$ will be considered composed of two parts i_C and $k_{\bar{C}}$ and j_C and $k_{\bar{C}}$ respectively where C is a set of variables within which <i> and <j> differs. With a loss function defined by:

$$\text{Loss}(a,b) = \begin{cases} 0 & \text{if } a=b \text{ or } n_a=0 \text{ or } n_b=0 \\ (n_a+n_b)\log_2(n_a+n_b) - n_a\log_2 n_a - n_b\log_2 n_b & \text{otherwise} \end{cases}$$

the loss in structure, expressed as a distance, becomes:

$$\begin{aligned} d_{ij} &= T(\text{before}) - T(\text{after merging i and j}) \\ &= \frac{1}{n}\left[\sum_{C} \sum_{k_{\bar{C}}} \text{Loss}(<i_C k_{\bar{C}}>, <j_C k_{\bar{C}}>) - \sum_{u=1}^{m} \text{Loss}(x_{iu}, x_{ju}) \right] \end{aligned}$$

This distance is small when objects are redundant, share many values, and values with respect to which they differ carry little information. It is large when the descriptions of objects represent significant differentiations within the m-dimensional space. Like the Mahalanobis distance, the above takes all objects into consideration that share some values with either of the two objects being compared. But, unlike the Mahalanobis distance, it takes account of the multi-ordinal nature of the distribution of objects and does not assume any ordering of values.

For binary attributes in all variables several simple distances have been used. To simplify the notation, let me represent the matching and mismatching of attributes associated with objects i and j in terms of a 2 by 2 contingency table:

i \ j	1	0	
1	a	b	e_i
0	c	d	$1-e_i$
	e_j	$1-e_j$	1

$$a = \frac{1}{m}\sum_{u=1}^{m} x_{iu}x_{ju}$$

$$e_i = \frac{1}{m}\sum_{u=1}^{m} x_{iu}$$

where a is the proportion of matching 1s between i and j, b is the proportion of mismatches with i's 0s cooccuring with j's 1s, etc. (the distance d_{ij} is not to be confused with the proportion d). In these terms, the Euclidian distance becomes:

$$d_{ij} = \sqrt{b + c}$$

and two simple matching coefficients used by Zubin (1938) and Jaccard (Socal and Sneath, 1963) respectively are:

$$d_{ij} = b + c$$

and:

$$d_{ij} = a + b + c$$

Sokal and Sneath (1963) also use Yule's association coefficient as a similarity measure:

$$S_{ij} = \frac{ad - bc}{ad + bc}.$$

Space does not permit to review and discuss the many distance and similarity measures possible and actually applied in clustering. The user of any clustering technique that starts from distance or similarity data must ascertain though that these measures possess the metric properties the clustering technique requires and that the assumptions implied by the choice of a particular distance measure conform to what the clusters are expected to represent.

Goals and Computational Efforts

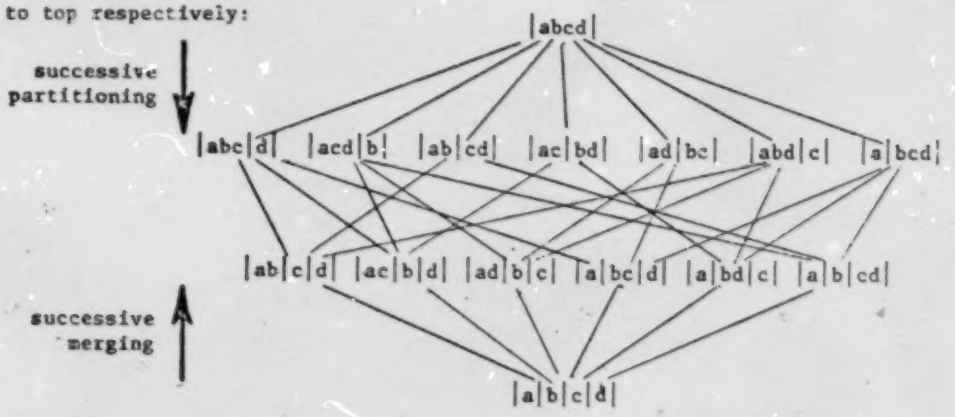
A major problem of all multivariate techniques is the amount of computation required to produce results. Since the number of cells in a many-dimentional space grows exponentially with the number of dimensions, such numbers often approach limits of computability before data can be considered rich enough to contain interesting information. Virtually all multivariate analysis algorithms rely on computational shortcuts to reduce this effort and thereby impose assumptions on the way data are processed. Cluster analysis is no exception and the user should know what is involved.

The task of clustering may be one of the following:
(a) selecting that subset of a set of objects which contains a designated object and of which all members satisfy specified criteria for inclusion into the set, class, type of cluster. (b) selecting that partition of a set of objects into a specified number of exhaustive and mutually exclusive subsets, classes, types or clusters, the parts of the partition, of which each is in a specificable sense optimal under the numerical restriction. (c) selecting that partition of a set of objects into exhaustive and mutually exclusive subsets which satisfies a specified criteria of optimality (without restriction on the number of parts in the partition). (d) selecting that binary decision tree which contains only partitions satisfying (b) and includes the partition satisfying (c).

The following table shows how the number of alternatives among which decisions need to be made grows with the number of objects to be clustered. With only n=10 objects, task (a) presents 512 alternatives which is a manageable and unproblematic number. But task (c) requires the evaluation of 116 thousand partitions, a number that approaches practical limits of computation while task (d) with its 2.6 billion decision trees already exceeds computational limits. The lesson to be drawn from any extension of numbers n of this table into the domain of theoretical significance is that practical clustering procedures cannot compute alternatives simultaneously but must proceed iteratively, in irreversible steps. It is in the form of iteration that hierarchical clustering schemes emerge.

n		1	2	3	4	5	6	7	8	9	10
(a) Number of subsets of which g is a member	2^{n-1}	1	2	4	8	16	32	64	128	256	512
(b) Number of partitions into k parts	$\sum_{j=0}^{k-1} (-1)^j \frac{(k-j)^n}{j!(k-j)!}$ k=1	1	1	1	1	1	1	1	1	1	1
	2		1	3	7	15	31	63	127	255	511
	3			1	6	25	90	301	966	3,025	9,330
	4				1	10	65	350	1,701	7,770	34,105
	5					1	15	140	1,050	6,951	42,525
	6						1	21	266	2,646	22,827
	7							1	28	462	5,880
	8								1	36	750
	9									1	45
	10										1
(c) Total number of partitions	$\sum_{k=1}^{n} \sum_{j=0}^{k-1} (-1)^j \frac{(k-j)^n}{j!(k-j)!}$	1	2	5	15	52	203	877	4,140	21,147	115,975
(d) Number of binary decision trees	$\frac{n!(n-1)!}{2^{n-1}}$	0	1	3	18	180	2,700	56,700	1,587,600	57,153,600	2,571,912,000

Two iterative clustering procedures can be distinguised: The successive partitioning of a set of objects into more and smaller subsets (classes or clusters) and the successive merging of objects into fewer and larger subsets (classes or clusters). Sneath and Sokal (1973) calls the former technique divisive and the latter agglomerative. With "|" separating the parts of a partition among four objects a, b, c, and d, these two options can be depicted as following a path through a partition lattice from top to bottom or from bottom to top respectively:



While either option results in the choice of one out of $n!(n-1)!2^{n-1}$ binary decision trees, their computational efforts are rather different as the following table of the number of alternatives may show:

	1st step	2nd step	...
Successive partitioning;	$2^{n-1}-1$	between $2^{(\frac{n}{2}-1)}-1$ and $2^{n-2}-1$	...
Successive merging:	$\frac{n(n-1)}{2}$	$\frac{(n-1)(n-2)}{2}$	...

Numerically, when n=100, the first step of successive partitioning poses about 10^{30} alternative partitions to chose from , a practical impossibility, whereas successive merging calls for the evaluation of only 4950. While successive partitioning cuts these numbers down rather quickly, evidently the first step is prohibitive.

Finally, computation is affected by two further distinctions. The first is due to Lance and Williams (1967) who define a combinatorial strategy as one in which the ori-

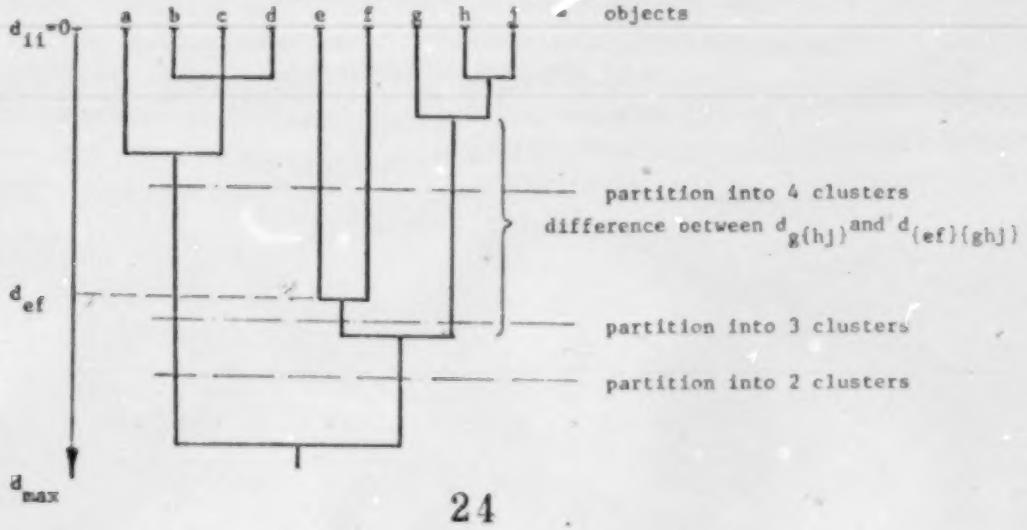
ginal input matrix (of distances or of data in canonical form) is successively transformed, becomes smaller and simpler and thereby reduces the computational effort by each step, whereas a non-combinational strategy is one in which all computations are based on the original input which must therefore be maintained throughout. Obviously combinational strategies are more efficient and among those one must distinguish between strategies that are either distance recursive when they derive each distance matrix from its preceeding distance matrix or data recursive when they successively modify the representation of data in canonical form. Distance recursive strategies are at least m-times (m=the number of variables) as efficient as data recursive methods.

Computational advantages may be outweighed, however, by the information that combinatorial strategies, and among these, distance recursive methods loose. Whether this can be justified is a question of validity.

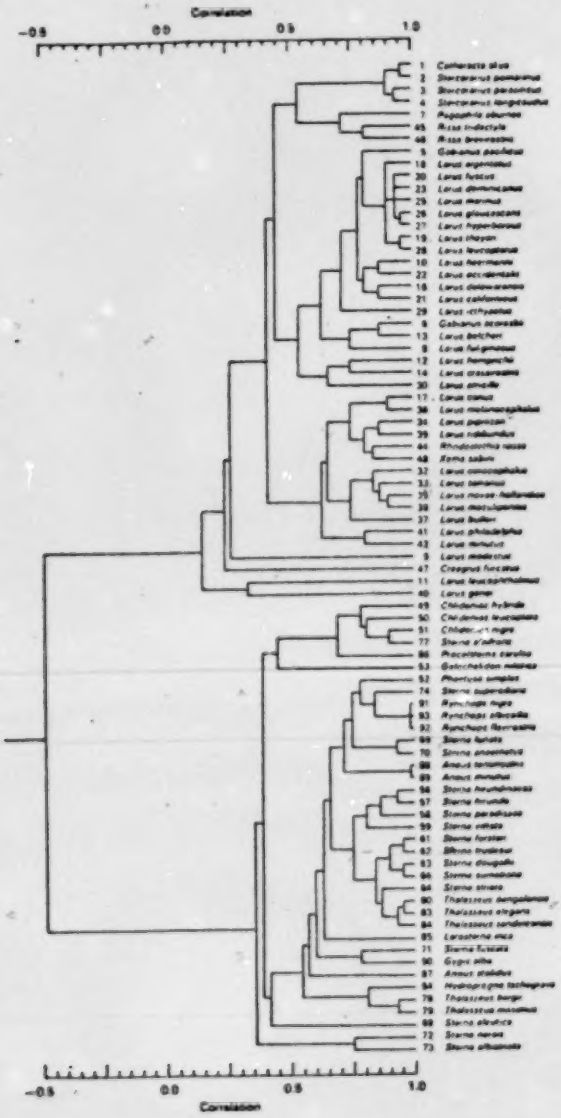
Presentation of Results

The conceptualization of multivariate data is difficult. We are just not accustomed to see point distributions in four or more dimensional spatial representations and it is this fact that often serves as a motivation for applying multivariate statistical analyses. All multivariate techniques transform such data. Even so analytical results may appear simple, it is in fact quite difficult to relate the transforms of these data to the original observations. This section does not deal with questions of meaningfulness and of the adequacy of the transformations for producing a result. It is rather concerned with several ways the results of a clustering process might be visualized leaving the conceptualization of the process for the subsequent section.

Dendrograms. The most important form of representing clustering results is the dendrogram which is a tree-like structure whose branches terminate at the objects being clustered and in which the lengths of its stems indicate differences in homogeneity, or heterogeneity within clusters being merged or partitioned. Dendrograms are nothing but a more sophisticated form of listing objects by their membership in clusters with each horizontal cut through a dendrogram indicating one of several partitions of the set of objects and the height between this cut and the original (unclustered) objects indicating the level of homogeneity or heterogeneity lost in this partition:



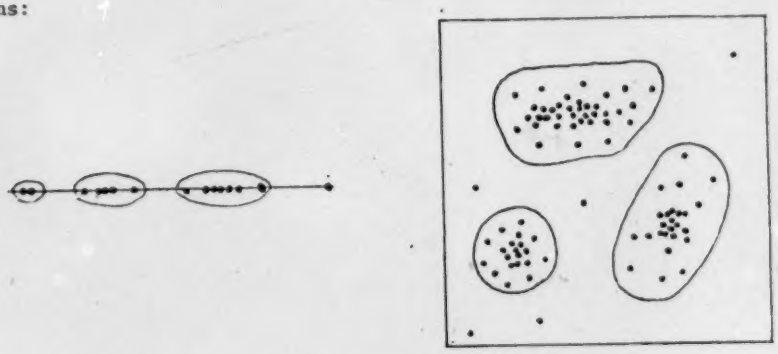
Dendrograms are particularly suited to represent a large number of objects and the whole history of a clustering process whether it proceeds by successive merging or partitioning. Relatively long stems between branches indicate to the researcher where relatively large jumps in heterogeneity occur and might suggest cut-off points at which partitions might be meaningful.



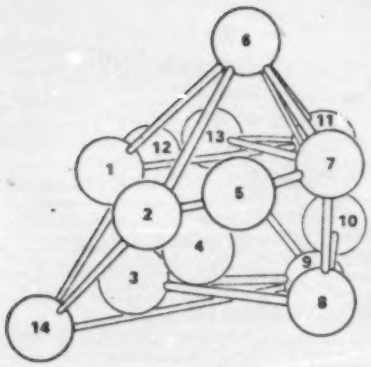
A modification of the dendrogram which is particularly suited for computer printouts is

used by Johnson (1967) and exemplified in the section on algorithms.

Spatial Representations for bi-ordinal clustering, by far the most appealing form of representation, depicts the set of objects together with distances between them in some space and indicates clusters by their boundaries. Since proximities are an essential ingredient of Gestalt perception, the conversion of similarities, correlation and the like into distances aids the visualization of groupings tremendously. For example, in one and two dimensions:

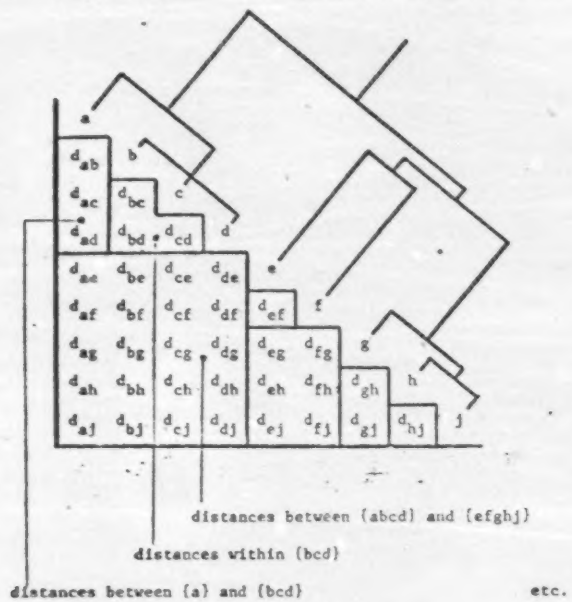


When three dimensions are involved the representation is somewhat more cumbersome although still possible. Spatial representations in four or more dimensions become virtually unreadable however. Since many multivariate data consist of objects that are characterized by many more than four variables, the use of spatial representations of clustering results is therefore extremely limited. But since two or three-dimensional representations are so common, researchers have resorted either to ignore the many-dimensional character of the distances between objects and approximate them by distances in two or three-dimensional representations or to employ dimension reducing techniques, such as factor analysis, that yield visually representable distributions of objects in a space with orthogonal dimensions. Since the size of the resulting clusters and distances between them are then distorted in the visual representation, numerical values for these may have to be added to indicate their true quantitative relationship. This is omitted in the following visualization of a taxonomy of the Enterobacteriaceae (Lysenko and Sneath, 1959):



It should be reemphasised that spatial representations with its emphasis on proximity carry strong bi-ordinal biases. Higher-order relations among three or more objects have no obvious spatial form.

Reordered Distance Matrices. Several authors, among them Sneath and Sokal (1973), suggest that the results of distance recursive merging be represented by reordering the entries of the initial distance matrix D^0 so that the proximity of rows (and columns) reflect the rank ordering of distances between objects. However, when replacing the requirement of ranking by the hiearchical ordering that any iterative clustering process imposes, one obtains a reordered distance matrix, such as the one below, in which entries are blocked into sections representing distances within and between clusters respectively.



While distances again bias the representation of clustering results, in favor of bi-ordinal conceptions, since all clustering procedures make use of some form of distance, there is no reason why multi-ordinal techniques should not employ such a form as well. Its drawback is that the motivation for clustering is then apparent only for one step. For an example see the output for the strong association clustering algorithm.

Typical Objects and Centroids. It is often desirable to identify an object, real or hypothetical, that is most representative of a cluster, called the typical object or centroid respectively. Centroids locate a given cluster in multivariate space. In monothetic clustering schemes, the centroid is that m-tuple of values that all objects in a cluster share. Since the objects in clusters emerging from polythetic techniques need not have any value in common, the centroid may then be quite abstract and is often unidentifiable in terms of data in canonical form. In either case, there may not exist a real object in the sample that coincides with the hypothetical centroid of a cluster. The object closest to the centroid of a cluster may then be chosen as the typical object of the cluster.

In single linkage clustering methods and those that employ measures of differences between clusters rather than within, centroids have no clear theoretical justification for clusters have this chain-like appearance and heterogeneities are not accounted for by these methods respectively. The computation of centroids and the identification of typical objects is thereby not excluded, however.

When data are available in canonical form, clustering proceeds from distances, and is, hence, bi-ordinal, the centroid of the cluster G becomes the m-tuple:

$$\langle \bar{x}_{G1}, \ldots, \bar{x}_{Gu}, \ldots, \bar{x}_{Gm} \rangle$$

where $\bar{x}_{Gu}$ may be interpreted as the arithmetic mean of values of the u-th variable over objects in cluster G when that variable has interval metric, as the median when that variable has ordinal metric, and as the mode when that variable is unordered.

Otherwise, the researcher tends to be restricted to identify the typical object of a cluster by:

$$\min_{g \in G}(d_{gG}) \quad \langle\text{=====}\rangle \quad g \text{ is the typical object of } G$$

in which d_{gG} stands for the heterogeneity measure chosen. In diametric clustering:

$$d_{gG} = \max_{j \in G}(d_{gj})$$

whereby the typical object g is the one closest to the center of the circle circumscribing the objects in G. In variance type clustering:

$$d_{gG} = \sqrt[r]{\frac{1}{n_G - 1} \sum_{j \in G} d^r_{gj}}$$

and when r=1, the typical object occupies a position close to the mean of the cluster. Large values for r make that position more responsive to the skewness of the distribution of objects within a cluster.

In multivariate classification the typical object of a cluster is defined as that object which, when removed from the cluster, causes the least amount of structure loss within the cluster. With $n_{\bar{g}u} = n_{Gu} - n_{gu}$ and $n_{\langle \bar{g}_C k_{\bar{C}} \rangle} = n_{\langle G_C k_{\bar{C}} \rangle} - n_{\langle g_C k_{\bar{C}} \rangle}$ for simplicity of notation:

$$d_{gG} = \frac{1}{n}\left[\sum_{C} \sum_{k_{\bar{C}}} \text{Loss}(\langle \bar{g}_C k_{\bar{C}} \rangle, \langle g_C k_{\bar{C}} \rangle) - \sum_{u=1}^{m} \text{Loss}(\bar{g}u, gu) \right]$$

If the proportion $\frac{n_{\langle g_C k_{\bar{C}} \rangle}}{n_{\langle G_C k_{\bar{C}} \rangle}}$ = constant for all objects that share some values with g, then $d_{gG}=0$ and g is both the typical object and centroid of the cluster G.

A drawback of representing clusters by typical objects is not only that there may exist objects that are actually more representative (closer to the centroid) of a cluster than those occurring in the sample but also that there may exist many different objects for which d_{gG} is equal and minimum. Ultimately, the first problem is one of sampling and the second one of measurement. Both can be overcome by permutation of the characteristics of possible objects. But this may not be a simple task.

Other Multivariate Techniques. Clustering, insofar as it provides a single partition of canonically represented objects in effect adds to the m descriptive values an index of each object's membership in some cluster. Such an index can be regarded as the m + 1st descriptive variable and expands the initial nxm data matrix in canonical form to an nxm+1 matrix. This expanded matrix may be subjected to a variety of other multivariate techniques, for example, multiple discriminant techniques yielding explanations of the clusters in terms of the features that discriminate.

The results of clustering can also be subjected to factor analysis, particularly yielding a system of coordinates for representing these clusters more efficiently (see Hunter in this volume). The results of different clustering techniques may be compared by crosstabulations, etc. There is no limit to the use of other analytical techniques for exploring the nature of the clusters that have been obtained as well as for the use of other techniques to provide or prepare the data for clustering. Thus other multivariate techniques may provide a host of other descriptive accounts of the results of clustering.

Properties of Emerging Clusters

The properties of objects within clusters, the boundaries around clusters, the relations between clusters, the tree-like dendrograms describing either the history of merging objects into classes or the history of partitioning sets of objects into subsets, all are expected to represent characteristics that are manifest in data. Decision criteria for clustering must therefore be based on measures that characterize sets of objects with distances being the special case in which these sets have exactly two members. Once the criteria for merging or for partitioning are set, the clusters that do emerge develop certain properties that should correspond to those manifest in data. This section presents several measures on clusters (as opposed to pairs) of objects and in terms of three dimensions of classification:

difference measures vs. heterogeneity measures
single linkage measures vs. multiple linkage measures
polythetic measures vs. monothetic measures

The distinction between bi-ordinal and multi-ordinal clustering is also reflected in these measures and the formal conditions imposed on such measures depend to a large part on the computational approach taken, i.e. whether the procedure is distance recursive or data recursive, and whether clusters are formed by partitioning (divisive) or by merging (agglomerative). Naturally the number of combinatorially possible clustering criteria exceeds those actually realized and of those available only a few can be discussed here.

Conceptually simple is the differentiation between difference measures and heterogeneity measures though the implications are far from obvious. Applied to clusters (classes of objects) difference measures quantitatively assess differences between two different clusters whereas heterogeneity measures assess differences within one cluster. Both are divergent generalizations of the distance between two objects. When a third object is added to a cluster of two, a difference measure assesses the difference between the third object and either one or both of the two objects already merged while a heterogeneity measure assesses some difference among all three objects regardless of how they were

brought together and thereby assigns equal weight to each object involved.

Notationally, d_{EF} will be used to denote the heterogeneity of the union of two sets of objects, whereas $d_{E|F}$ will be used to denote the difference between the two sets. The distance d_{ij} between two objects i and j then is the special and overlapping case at which E=(i) and F=(j).

In these terms, and without reference to details, several formal requirements on the use of these measures as decision criteria for clustering can be stated. For distance recursive procedures, both heterogeneity and difference measures must satisfy analogous conditions:

$d_{EE} \geq 0$		
$d_{EF} \geq d_{EE}$	$d_{E\|F} \geq 0$	positive
$d_{EF} = d_{FE}$	$d_{E\|F} = d_{F\|E}$	symmetrical
$d_{EF} \leq d_{EG} + d_{GF}$	$d_{E\|F} \leq d_{E\|G} + d_{G\|F}$	triangle inequality

to which may have to be added:

$d_{EF} \leq \max(d_{EG}, d_{GF})$	$d_{E\|F} \leq \max(d_{E\|G}, d_{G\|F})$	ultrametric inequality

These conditions correspond to those for distances between two objects except that the heterogeneity within a cluster, d_{EE}, may exceed zero and that the difference between a cluster and itself, $d_{E|E}$, is meaningless by definition.

For data recursive procedures these formal requirements may be relaxed to the following two conditions:

(1) Decision criteria must not decrease with each upward move in the partition lattice (merging) and not increase with each downward move in the partition lattice (partitioning). For any three clusters E, F, and G, at any stage in the clustering process:

$$\max(d_{GG}) \leq \min(d_{EF})$$

For difference measures this does not apply generally. However, if one interpretes $d_{G|G}$ as that distance which served as criteria to merge two clusters

into G, then, while some differences within G may exceed $d_{E|F}$, at least under a successive merging strategy:

$$\max(d_{G|G}) \leq \min(d_{E|F})$$

This condition favors heterogeneity measures because whenever G is the union of E and F, $d_{GG} = d_{EF}$. For difference measures the condition must be rephased to read: the largest of the differences that lead to the formation of a cluster must not exceed the smallest difference between any pair of clusters at the time at which that cluster was formed.

(2) Ideally, decision criteria should also be independent of the order of cluster formation: this is again satisfied by heterogeneity measures:

$$d_{\{ijk\}\{ijk\}} = d_{\{i\}\{jk\}} = d_{\{ij\}\{k\}} = d_{\{ik\}\{j\}}$$

but not by difference measures. Some of the implications of this failure will become apparent below.

The second classificatory dimension refers to whether clusters are characterized by a single representative linkage between two objects or whether the measure aggregates multiple linkages into a single index. I will exemplify the two dimensions by several individual and polythetic measures for bi-ordinal clusters.

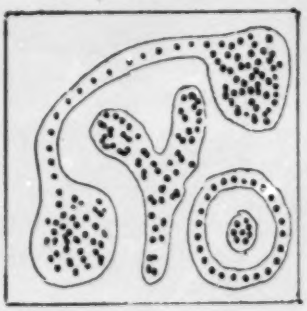
	difference	heterogeneity
single linkage	connectedness	diameter
multiple linkage	average linkage	variance

Clusters that are characterised by the connectedness of their members (Johnson, 1967), single linkage clustering (Sokal and Sneath, 1963) stem from the simplest form of clustering with difference measures: Clusters are formed by merging objects in the increasing order of their distances, i.e. by merging those clusters that contain at least one object each which is least distant across cluster boundaries. Thus, the two clusters E and F, if they are to be merged to form the cluster G satisfy the criterion:

$$\min_{E,F}(d_{E|F}) = \min_{E,F}\left(\min_{\substack{i\in E \\ j\in F}}(d_{ij})\right)$$

The difference $d_{E|F}$ in this criterion is crucially dependent on the history of the formation of the cluster. It is the distance between two objects within G that, at the point of the last addition to the cluster, was the smallest distance between objects of all different clusters formed prior to G. Within a cluster it is possible to find distances that are larger or smaller then the distance on account of which the cluster was formed.

While leading to an extremely simple form of computation, (in fact it can be done manually by inspection of the distance matrix D without any arithmetic computation) the weakness of this criterion lies in its tendency to form long chains that often bridge otherwise perfectly meaningful clusters. Some of the peculiar clusters that this criterion will produce are illustrated in the following spatial representation. The use of difference measures generally and in conjunction with single linkage conceptions of those measures in particular, allows clusters to "grow out of control;"



Clusters may be characterized by the largest distance between their members, called their diameter. The technique minimizing the diameter of a cluster is variably called the complete linkage method (Sokal and Sneath, 1963), compact clustering (Lorr, 1968) the diameter method (Johnson, 1967) and controls for what single-linkage clustering omits, that extreme differences within a cluster, stay within bounds. The largest distance within a cluster G, the diameter is:

$$d_{GG} = \max_{i,j\in G}(d_{ij})$$

The diameter of a cluster is a heterogeneity measure but it takes only one distance as representative of the cluster as a whole. In the formation of clusters that minimize this measure, a successive merging procedure will join at each step those two clusters whose most distant objects have the smallest distance across all pairs of clusters. Thus, by analogy to MacNaughton-Smith (1965) and Johnson (1967), if E and F are merged to form a new cluster G:

$$d_{CG} = \min_{E,F}(d_{EF}) = \min_{E,F}\left(\max_{\substack{i\in E \\ j\in F}}(d_{ij})\right)$$

Compared with the connectedness criterion, the diameter criterion yields relatively compact clusters. But such clusters have several peculiar properties. First, there is the tendency of clusters to become circular and equal in diameter. Second, the number of objects within a cluster, the density of its population, has no bearing on the way clusters are formed. Third, and a corollary of the second, a center need not exist for such clusters. To illustrate some of the typical clusters diametric clustering will yield, one may examine the following distribution of objects with clusters shown at two stages of formation.

Average linkage clustering (Sokal and Sneath, 1963) extends the notion of connectedness to an aggregate measure of all distances between two clusters. So, Bock (1974) defines the average distance between two clusters by:

$$d_{E|F} = \frac{1}{n_E n_F}\sum_{i\in E}\sum_{j\in F} d_{ij}$$

The criterion to merge E and F into G, generalized to any power, then is:

$$\min_{E,F}(d_{E|F}) = \min_{E,F}\left(\sqrt[r]{\frac{1}{n_E n_F}\sum_{i\in E}\sum_{j\in F} d_{ij}^{r}}\right)$$

Another average linkage criterion is Pearson's(1926) coefficient of radical likeness:

$$\min_{E,F}(d_{E|F}) = \min_{E,F}\left(\sqrt{\frac{1}{m}\sum_{u=1}^{m}\frac{(\bar{x}_{Eu}-\bar{x}_{Fu})^2}{\frac{\sigma_{Eu}}{n_E}+\frac{\sigma_{Fu}}{n_F}}}\right)$$

in which $\bar{x}_{Eu}$ is the arithmetic mean of values in E of variable u, and σ_{Eu} is the variance within E regarding the uth variable. The coefficient has been used by Rao (1948, 1952), Sokal and Sneath (1963) and by several others. The coefficient makes the difference between two clusters relative to the variance within both and is thus not a "pure" average linkage measure but belongs to the same family.

Both difference measures work against the chainlike appearance of clusters typical of the connectedness criterion but do not eliminate it completely. Their clusters are less compact than those using corresponding heterogeneity measures. Difference measures simply do not optimize homogeneity within a cluster. Once two clusters are merged, the distances that contributed to that decision are no longer referred to in subsequent clustering steps.

Clusters with minimum variance between its objects are achieved by taking all distances within a prospective cluster into account. Accordingly, a multiple linkage measure of heterogeneity within a cluster G is:

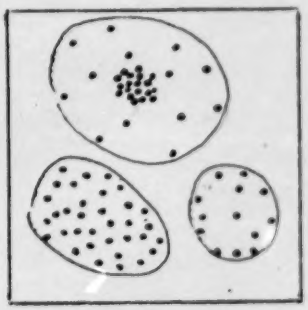
$$d_{GG} = \sqrt[r]{\frac{1}{n_G(n_G-1)}\sum_{i,j\in G} d_{ij}^r}$$

and to minimize this heterogeneity, clusters E and F are merged into G when:

$$d_{GG} = \min_{E,F}(d_{EF}) = \min_{E,F}\left(\sqrt[r]{\frac{1}{(n_E+n_F)(n_E+n_F-1)}\sum_{i,j\in E,F} d_{ij}^r}\right)$$

When the exponent r=1, d_{GG} is the mean distance within the objects of a cluster and its use as a decision criterion assures that this mean distance is kept at a minimum. When r=2, d_{GG} is the standard deviation within a cluster. And, since the variance is the square of the standard deviation, clustering with r=2 also might be said to minimize the variance within clusters. Sokal and Sneath (1963) termed the latter index, the taxionomic distance.

Variance type heterogeneity measures compensate a large distance within a cluster by several smaller distances within that cluster. Clusters with equal heterogeneity in this variance sense may thus be different in diameter. However, as r increases in value, larger distances weigh increasingly heavy on the measure so that r in fact controls the conservatism of the clustering procedure. The higher the exponent r the more compact the clusters that emerge. Some typical clusters that variance methods will identify are depicted as follows:



The relationship between variance type difference and heterogeneity measures is easily illustrated by the following equality in which r=1 for both the average linkage between clusters and the mean distance within the union of two clusters:

$$(n_E+n_F)(n_E+n_F-1)\ d_{EF} = n_E(n_E-1)\ d_{EE} + n_E n_F\ d_{E|F} + n_F(n_F-1)\ d_{FF}$$

This equality reveals the difference measure $d_{E|F}$ to be only one part of the heterogeneity measure d_{EF}. If used as a clustering criterion, $d_{E|F}$ ignores the heterogeneities d_{EE} and d_{FF} of the clusters being merged and thus minimizes some property other than a characteristic of all the objects in the emerging cluster. This accounts for what I suggested earlier that difference measures tend to let clusters"grow out of control" and are for this reason inferior to measures of heterogeneity.

Clusters may also be polythetic or monothetic (Sokal and Sneath 1963). In a monothetic cluster all members share some properties. Objects are admitted to a cluster because a large number of their characteristics match. In a polythetic cluster, members need not hold any value in common. They need only be similar in some respect and this similarity may be expressed by a high correlation, by a larger proximity between values or by the sharing of values between pairs rather than among all members of a cluster. Distance re-

cursive techniques yield polythetic clusters only and all clustering criteria discussed so far are also polythetic in result.

An example of a monothetic clustering technique is Krippendorff's (1975) strong associative clustering of binary attribute data. Key to the technique is the successive enumeration of the attributes that are shared among all members of a cluster. Since this number cannot be obtained from agreements between pairs, the enumeration must be data recursive. And, seeking to correct observed agreement on attributes by what is due to chance, one of the more convincing coefficients turns out to be a generalization of Benini's (1901) measure of association. With a_G as the proportion of attributes shared among objects in G and e_i as the proportion of attributes associated with the object i of G, the coefficient, converted into a heterogenity measure by $d_{ij}=1-s_{ij}$, and stated as a decision criterion for merging E and F into G is:

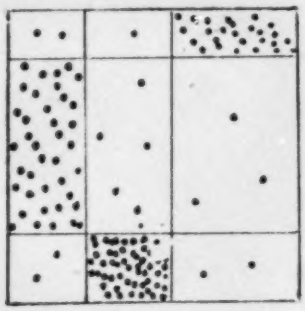
$$d_{GG} = \frac{\min_{i\in G}(e_i) - a_G}{\min_{i\in G}(e_i) - \prod_{i\in G} e_i} = \min_{E,F}(d_{EF}) = \min_{E,F}\left(\frac{\min_{i\in E,F}(e_i) - a_{EF}}{\min_{i\in E,F}(e_i) - \prod_{i\in E,F} e_i}\right)$$

It is the proportion of the observed disagreement on attributes shared within G to the disagreement of chance matching.

A graphical representation of typical clusters resulting from this monothetic heterogeneity measure is difficult precisely because multi-ordinal clusters defy spatial representations. Further elaboration of the measure and an example are found in the section on clustering algorithms. This heterogeneity measure serves as an example of a measure that does not satisfy the triangle inequality and would thus not lend itself to distance recursive clustering.

An example of a polythetic multi-ordinal and variance type technique is multivariate classification. It is a method by which clusters are formed neither by grouping objects in terms of their variables nor by grouping variables in terms of their objects (all of which might produce one univariate classification or clustering scheme) but by clustering variables in terms of each other, interactionally so to speak, using the distribution of

objects in multivariate space as a reference for the interaction. The measure used by this technique is one of several developed from information theory by Krippendorff (1969, 1974) of which its simplest form is given in the section on distances. For simplicity, its recursive form is presented in the section on algorithms where details of the measure are elaborated. This measure assesses the amount of multivariate structure lost within the m-dimensional space when some of the terms within variability are no longer differentiated. It can be interpreted as expressing the amount of multivariate information that can no longer be transmitted due to the formation of clusters in each dimension or as the amount of relational entropy lost within a cross classification of m separate clustering schemes, one for each variable. What this clustering technique achieves is a more efficient representation of the objects involved, one that reduces the m-dimensional space in volume without much loss in the essential relationship. Since the complexity of the resulting cluster again defies a simple graphical representation, I offer a two dimensional diagram of the nature of the clusters that the technique might identify.



In conclusion, it is evident that the choice of clustering criteria is the most important determinant of the kind of clusters that do emerge and the kind of properties these clusters are thereby able to represent. The classification of these properties merely serves to make differences between clustering criteria and the emerging properties of clusters transparent. Examples are more numerous for difference than for heterogeneity measures, for single than for multiple linkage procedures, for polythetic than for monothetic clusters. The reliance on bi-ordinal rather than multi-ordinal conceptions of properties is striking and the fact that I did not exemplify the results of partitioning

approaches to clustering is indicative of suspicious white spaces on the map of all combinatorially possible clustering techniques.

Clustering Algorithms

An algorithm is a stepwise procedure that is completely specified (leaves no alternative undecidable) and transforms some input into some output. Clustering algorithms accept data as input either in their canonical form of a data matrix X or in their derived form of a distance matrix D. The output of a clustering algorithm either identifies a cluster containing a given object, produces some partition, a set of clusters satisfying some criterion of optimality, or it gives a decision tree that contains partitions all of which satisfy some criterion of optimality.

Because of the insurmountable efforts required to compute clusters, partitions and decision trees simultaneously, demonstrated above, algorithms must be defined recursively. A recursive algorithm is one that is applied to some initial set of data, yielding an output to which it is applied again and again until some terminating criterion is met. Recursive clustering algorithms work themselves stepwise either down a partition lattice (through successive partitioning) or up the partition lattice (through successive merging). After the first step, a recursive algorithm avoids references to the initial data and transforms only its transforms.

This section presents four different algorithms for clustering objects or variables. The algorithms are chosen for their distinctive features: successive partitioning vs. successive merging, bi-ordinal clustering vs. multi-ordinal clustering, data recursive vs. distance recursive, monothetic vs. polythetic , etc. The algorithms presented here do not exhaust all alternatives, however. A researcher has many more options than are given here.

The Johnson-algorithm. The first example of a clustering algorithm is taken from Johnson (1967) who formalizes the two single-linkage clustering techniques discussed above. Since both are distance-recursive merging techniques and proceed identically except for their clustering criteria, I will describe here only one, the algorithm for diametric clustering. The Johnson-algorithm is extremely simple, speedy in execution and therefore inexpensive to use which might explain its widespread application.

Given: The n×n matrix of distances $(d_{ij}^s) = D^s$ at the initial s=0.

Step 1: Search for the smallest distance, $\min_{E \neq F}(d_{EF}^s)$ in the (n-s)×(n-s) matrix D^s.

Step 2: Print or store on a history record: s+1, E, F for which d_{EF}^s is minimum and merge all pairs of objects in E and F into G.

Step 3: Compute a new (n-s-1)×(n-s-1) matrix of distances in D^{s+1} from D^s by replacing all entries with references to E and F by:

$$d_{GI}^{s+1} = \max(d_{EI}^s, d_{IF}^s) \text{ for all } I \neq E,F \text{ in } D^s$$

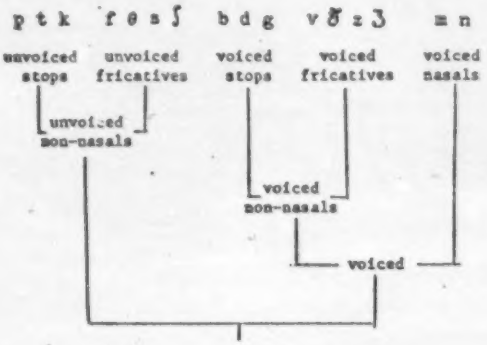
Step 4: Set s←s+1, return the new D^s to step 1 unless either the smallest distance d_{EF}^{s-1} now d_{GG}^s exceeds a specified limit , the number of remaining clusters n-s falls below a specified number, or any other terminating criterion is satisfied.

Step 5: Print results and terminate.

For an illustrative example, Johnson uses data obtained from a psychoacoustic study of sixteen principle consonants which are listed here across the table. The numbers down the left hand side are similarity values that were obtained in the study and are here associated with each merging as indicated.

```
                p t k f θ s ʃ b d g v ð z ʒ m n
 - - -          . . . . . . . . . . . . . . . .
2.635           . . . . . . . . XXX . . . . . .
2.234           XXX . . . . . . XXX . . . . . .
2.230           XXX . . . XXX . XXX . . . . . .
2.123           XXX . . . XXX . XXX . . XXX . .
1.855           XXXXX . . XXX . XXX . . XXX . .
1.683           XXXXX . XXXXX . XXX . . XXX . .
1.604           XXXXX . XXXXX . XXX . . XXX XXX
1.525           XXXXX . XXXXX . XXX . XXXXX XXX
1.186           XXXXX . XXXXX . XXX XXXXXXX XXX
1.119           XXXXX . XXXXX XXXXX XXXXXXX XXX
0.939           XXXXX XXXXXXX XXXXX XXXXXXX XXX
0.422           XXXXXXXXXXXXX XXXXX XXXXXXX XXX
0.302           XXXXXXXXXXXXX XXXXXXXXXXXXX XXX
0.019           XXXXXXXXXXXXX XXXXXXXXXXXXXXXXX
0.000           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Apparently the resulting clusters correspond to the distinctive features presumed by Miller and Nicely (1955) who provided these data. At the level of five clusters the 16 phonemes divide into the following hierarchy:



The CONCOR-algorithm. The second example of a clustering algorithm is taken from Breiger, Boorman and Arabie (1975). It is the only effective partitioning algorithm I know that is applicable to a wide variety of data. CONCOR is the acronym of "convergent correlation" which designates the iterative application of correlations to yield stable indications of dependencies which are in turn used to partition a set of objects or variables into two parts. The use of correlations identifies the method as a bi-ordinal partition technique.

Given: The $n \cdot m$ matrix $(x_{iu}) = X$ with interval metric in each variable u

Step 1: Compute the $n_G \cdot n_G$ matrix of product-moment correlations ${}_o(r^s_{ij})_G = {}_oR^s_G$ for the initial s=0 where the initial set G is the set of all $n_G = n$ objects.

Step 2: Compute the $n_G \cdot n_G$ matrix of product-moment correlations ${}_{t+1}(r^s_{ij})_G = {}_{t+1}R^s_G$ from all pairs i and j in ${}_tR^s_G$ and iterate this step to create from ${}_oR^o_G = {}_oR^s_G$, ${}_1R^s_G$, ${}_2R^s_G$, ${}_3R^s_G$, ... until all entries r^s_{ij} of ${}_wR^s_G$ approximate within a specified limit the value +1 or -1.

Step 3: Permute ${}_wR^s_G$ into the bipartite form:

	E	F
E	+1	-1
F	-1	+1

Step 4: Print or store on a history record: s+1, the partition of G into E and F, and decompose the original correlation matrix ${}_oR_G^o$ into and store separately the two submatrices, the $n_E \cdot n_E$ matrix ${}_oR_E^o$ and the $n_F \cdot n_F$ matrix ${}_oR_F^o$.

Step 5: Set s←s+1, search for the largest remaining matrix ${}_oR_I^o$ and return that matrix as ${}_oR_G^s$ to step 2 unless either s exceeds a specified limit or the number n_G of the largest cluster falls below a specified number.

Step 6: Print results and terminate.

The authors developed this algorithm for clustering a variety of sociometric data and profiles of either objects or variables between which correlations can be computed. This requires that all variables possess interval metric. There is no reason, however, to restrict the use of this algorithm to canonical forms of data and to variables with interval metrics. Since the principal feature of the algorithm is that correlations are computed iteratively, any matrix of distances or similarities with appropriate interval metric properties could be entered at step 1 in place of ${}_oR_G^o$.

The authors have applied this algorithm to many sets of data and report that with n=70 and the cut-off point for $r_{ij} = .999$, no more than 11 iterations are needed to approximate stability conditions. This keeps the required computational effort within practical limits. Apparently, while there are a few theoretical examples of a knife-edge character in which the iteration of correlations does not converge to the bipartite form, actual data that would lead to indecisions of this sort have not been encountered.

The interpretation of the CONCOR clusters is difficult, however. As the authors recognise, the procedure does not use measures of homogeneity or heterogeneity as decision criteria and while it is easy to understand when correlations are positive within clusters and negative across, such an understanding is indeed difficult when one is concerned with correlations of correlations of correlations ... that might be 11 times removed from the data. Nevertheless Breiger, Boorman and Arabie have compared CONCOR results with a variety of results obtained from other clustering techniques and found them convincing. Its clus-

ters seem to be very similar to those obtained by clustering techniques based on a connectedness criterion.

The Strong Association-algorithm. In the next example I shall attempt to illustrate both how a very simple multi-ordinal clustering procedure works and how the required decision criteria may be formulated recursively to keep computational efforts small. It will be recalled that the computational effort of multi-ordinal clustering is generally magnified by the fact that such techniques require constant interaction between the procedure and data in their canonical form. Since clustering, to be practical, must proceed recursively, measures that keep account of the increasing heterogeneity in the emerging clusters should be defined recursively as well, else the procedure would have to return to the original data at each step and threby anull the computational advantage of recursion.

The algorithm for strong association of 2^m data (Krippendorff, 1975) was developed in this way. It is applicable to binary attribute data where the attributes to be shared within clusters are assigned the value $x_{iu}=1$ and the absence of this attribute the value $x_{iu}=0$. In terms of the 2 by 2 contingency matrix defined earlier, Krippendorff's generalization of Benini's association coefficient, which is converted here to a suitable heterogeneity measure by $d_{ij}=1-s_{ij}$, is:

$$d_{EF} = \frac{\min_{i\varepsilon E,F}(e_i) - a_{EF}}{\min_{i\varepsilon E,F}(e_i) - \prod_{i\varepsilon E,F} e_i}$$

where, in terms of the canonical form of data, $e_i = \frac{1}{m}\sum_{u=1}^{m} x_{iu}$ is the proportion of attributes present in object i, $a_{EF} = \frac{1}{m}\sum_{u=1}^{m} \prod_{i\varepsilon E,F} x_{iu}$ is the proportion of attributes shared within E and F, $\min_{i\varepsilon E,F}(e_i)$ is the largest possible proportion of attributes that could be shared within E and F, and $\prod_{i\varepsilon E,F} e_i$ is the expected proportion of attributes that would be shared if their cooccurence were due to chance.

It turns out, all these components of the measure can be defined in terms of one matrix and two quantities for each cluster at the initial s=0, and at any s+1 from s. At s=0 and from the initial n*m matrix X^0 with entries x_{iu}, the maximum proportion of attributes in cluster {i}, containing just one member, is:

$$\mu^{o}_{(i)} = \frac{1}{m}\sum_{u=1}^{m} x_{iu} = e_i$$

and so is the initial probability of attributes in that one-object cluster:

$$\rho^{o}_{(i)} = \frac{1}{m}\sum_{u=1}^{m} x_{iu} = e_i$$

At each prospective merger of two clusters E and F into G, these quantities change as follows:

$$\mu_G^{s+1} = \min(\mu_E^s, \mu_F^s)$$
$$\rho_G^{s+1} = \rho_E^s \rho_F^s$$

And the $(r-s)\times m$ matrix X^s becomes the $(n-s-1)\times m$ matrix X^{s+1} by:

$$x_{Gu}^{s+1} = x_{Eu}^s x_{Fu}^s$$

So that the measure of heterogeneity for merging E and F into G becomes a function of values solely available at the preceeding iteration:

$$d_{GG}^{s+1} = d_{EF}^s = \frac{\min(\mu_E^s,\mu_F^s) - \frac{1}{m}\sum_{u=1}^{m} x_{Eu}^s x_{Fu}^s}{\min(\mu_E^s,\mu_F^s) - \rho_E^s \rho_F^s}$$

This recursive formulation provides the key to the following surprisingly efficient algorithm:

Given: The $n\times m$ matrix $(x_{iu}^s) = X^s$ at $s=0$ with the presence of an attribute denoted by $x_{iu}=1$ and its absence by $x_{iu}=0$.

Step 1: For $i=1, 2, \ldots, n$ set $\mu^s_{(i)} = \rho^s_{(i)} = \frac{1}{m}\sum_{u=1}^{m} x_{iu}$

Step 2: Compute the $(n-s)\times(n-s)$ distance matrix $(d_{EF}^s)^s = D^s$ from the $(n-s)\times m$ matrix X^s by:

$$d_{EF}^s = \frac{\min(\mu_E^s,\mu_F^s) - \frac{1}{m}\prod_{u=1}^{m} x_{Eu}^s x_{Fu}^s}{\min(\mu_E^s,\mu_F^s) - \rho_E^s \rho_F^s}$$

Step 3: Search for $\min(d_{EF}^s)$ in D^s and for this smallest distance, print or store on a history record: s+1, $\min(d_{EF}^s)$, E, F, the newly assigned label G and D^s if required.

Step 4: Compute those recursive accounts that are affected by the merger of E and F into G:

$$\mu_G^{s+1} = \min(\mu_E^s,\mu_F^s)$$
$$\rho_G^{s+1} = \rho_E^s \rho_F^s$$

$$x^{s+1}_{C_u} = x^{s}_{Eu}\ x^{s}_{Fu}$$
$$s \leftarrow s+1$$

Step 5: Return the reduced (n-s)×m matrix X^s to step 2 unless either $\max(d^s_{GG})$ exceeds a specified value or any other terminating criterion is satisfied.

Step 6: Print result and terminate.

The example given here starts with an initial data matrix X^o describing 10 objects in terms of 16 variables. When the distribution of attributes are examined in this matrix, one may discover that the attributes of object j are fully contained in the attributes of the i-th object, yielding, as it should, a distance of $d^o_{ij}=0$. Since monothetic-clusters are represented by the attributes its objects share, the cluster {i,j} then takes on the attributes i and j have in common, here those of j, which may be seen in the subsequent transform of the data matrix. Also objects e and f show the strongest possible association which $d^o_{ef}=0$ indicates. At the third iteration it is the objects g and h that are found least distant. In terms of the 2 by 2 contingency table, the distance would be computed as follows:

h \ g	0	1	
0	$\frac{6}{16}$	$\frac{2}{16}$	
1	$\frac{1}{16}$	$\frac{7}{16}$	$\frac{8}{16}$
	$\frac{9}{16}$		1

$$d_{gh} = \frac{\min(\frac{8}{16},\frac{9}{16}) - \frac{7}{16}}{\min(\frac{8}{16},\frac{9}{16}) - \frac{8}{16}\frac{9}{16}} = .29$$

d_{gh} of D^2 then appears in D^3's diagonale as associated with the cluster {g,h}. So the process continues as indicated in the following dendrogram:

	X^s	D^s							
a	0 1 1 1 0 1 1 1 0 1 1 1 1 0 1 0	.00							
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.96	.00						
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	1.16	.64	.00			s = 1		
d	0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 0	.71	.89	1.07	.00				
e	0 0 1 1 1 1 1 1 1 1 1 0 0 1 1 0	.83	.96	.87	.00	.00			
f	0 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1	.36	1.19	1.42	.76	1.07	.00		
g	1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 1	1.20	1.00	1.60	1.14	1.60	.57	.00	
h	1 1 0 0 1 0 1 0 0 0 1 1 1 1 1 0 1	1.42	1.19	1.42	1.27	1.78	1.02	.29	.00
i	0 1 0 1 1 1 1 0 1 0 0 1 0 1 1 1	1.28	1.60	1.28	1.19	.96	.89	.67	.89	.00	
j	0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 0	.91	1.91	1.81	.98	.91	.98	.86	.65	.00	.00

a	0 1 1 1 0 1 1 1 0 1 1 1 1 0 1 0	.00								
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.96	.00							
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	1.16	1.00	.00			s = 3			
de	0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 0	.42	.68	.63	.00					
f	0 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1	.36	1.19	1.42	.54	.00				
g	1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 1	1.20	1.00	1.60	.82	.57	.00			
h	1 1 0 0 1 0 1 0 0 0 1 1 1 1 1 0 1	1.42	1.19	1.42	.91	1.02	.29	.00		
ij	0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 0	.50	1.17	1.00	.56	.66	.62	.44	.00	

a	0 1 1 1 0 1 1 1 0 1 1 1 1 0 1 0	.00								
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.96	.00							
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	1.16	.64	.00			s = 4			
de	0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 0	.42	.68	.63	.00					
f	0 1 1 0 0 1 1 0 1 1 1 1 0 0 1 1	.36	1.19	1.42	.54	.00				
gh	1 1 0 0 1 0 1 0 0 0 1 1 0 0 0 1	.82	.77	1.02	.80	.55	.29			
ij	0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 0	.50	1.17	1.00	.56	.66	.52	.00		

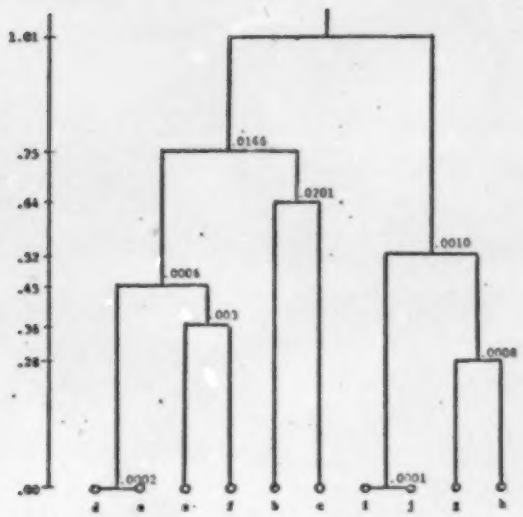
af	0 1 1 0 0 1 1 0 0 1 1 1 0 0 1 0	.36								
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.97	.00							
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	1.05	.64	.00			s = 5			
de	0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 0	.45	.68	.63	.00					
gh	1 1 0 0 1 0 1 0 0 0 1 1 0 0 0 1	.64	.77	1.02	.80	.29				
ij	0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 0	.56	1.17	1.00	.56	.52	.00			

adef	0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 0	.45								
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.67	.00							
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	.68	.64	.00			s = 6			
gh	1 1 0 0 1 0 1 0 0 0 1 1 0 0 0 1	.92	.77	1.02	.29					
ij	0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 0	.79	1.17	1.00	.52	.00				

adef	0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 0	.45								
b	1 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1	.67	.00							
c	1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1	.68	.64	.00			s = 7			
ghij	0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0	.88	.96	.98	.52					

adef	0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 0	.45								
bc	1 0 1 1 0 1 0 1 0 1 0 0 1 0 0 1	.75	.64							
ghij	0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0	.88	1.08	.52			s = 8			

abcdef	0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0	.75								
ghij	0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0	1.01	.52				s = 9			



The researcher now has to decide which partition is the most meaningful one, how clusters can be interpreted. This is in part an intuitive decision but it can be strengthened by statistical considerations. In this example the null-hypothesis (that the sharing of attributes within clusters is due to chance) can be rejected on significance levels indicated in the dendrogram at each point of merger. One can see that the significance level drops sharply after the sixth step and one might on the grounds be lead to accept the partition of the ten objects into four clusters as optimal. However, the attributes then still overlap. A perfect differentiation between clusters is achieved only before the last step at the point of which the two remaining clusters share no attribute any more, with 7 out of the 16 variables providing the basis of the differentiation. All others dropped out.

The Multivariate Classification-algorithm. This example presents a clustering technique that does not cluster objects for their own sake, rather uses them as a vehicle for simplifying their multivariate description. The description of these objects is a qualitative one, i.e. variables have nominal metric throughout. Given a many-dimensional distribution of objects, the task of the clustering procedure is to reduce the representational space of this distribution not in dimensionality but in size without or with only small amount of losses in structure within this space. Clusters then emerge not in one variable (e.g. the set of objects) and in terms of all other variables, rather, clusters emerge in

all variables simultaneously, each in terms of all others. What is thereby taken account of is that the clustering within one variable may interact with the clustering in another variable and that higher-order dependencies within data maybe the cause of such interactions. The most distinctive feature of this algorithm is that it optimises the representation of multi-ordinal relations in data by the simultaneous clustering of values within many variables.

I am presenting a version here of the algorithm that constitutes a considerable simplification over the one initially published (Krippendorff, 1974) and although multivariate classification provides several choices of heterogeneity measures, only the information theoretical measure of the amount of loss in structure will be used in the example. Since the procedure is a multi-ordinal one, and hence proceeds data recursively the recursive formulation of loss functions is a key factor to the algorithm's practicality.

The algorithm yields several, as many as there are variables, hierarchical clustering schemes for the qualities in terms of which the objects are described.

Given: The n×m matrix $(x_{iu}) = X$, all variables with nominal metric (unordered values).

Step 1: Reduce the n m matrix X to an (n-s)×m matrix X^s containing n-s unique objects i to each of which is assigned the frequency $n_{\{i\}}$. Compute frequencies $n_{x_{.u}} = \sum_{i=1}^{n-s} n_{x_{iu}}$ for each value x occuring in variable u. (From here on X^s serves as a matrix of indices only). Set $d_{\{i\}\{i\}}=0$ for all i=1, 2, ..., n-s.

Step 2: With the function: $\text{Loss}(a,b) = \begin{cases} 0 & \text{iff } a=b, \text{ or } n_a=0 \text{ or } n_b=0 \\ (n_a+n_b)\log_2(n_a+n_b) - n_a\log_2 n_a - n_b\log_2 n_b & \text{otherwise,} \end{cases}$

with $n_{x_{Eu}}$ denoting the frequency of the value x_{Eu} within variable u in terms of which objects in cluster E are characterized, and with the m-valued description of each object divided into two parts, E_C and $K_{\bar{C}}$, so that $n_{<E_C K_{\bar{C}}>}$ denotes the number of objects that share values x_{ic} within the set C of variables with objects in E but differ with respect to the remaining variables $\bar{C}$, now then

compute the new (n-s)x(n-s) distance matrix D^s, replacing missing distances only, by:

$$d^s_{EF}=d^s_{EE}+d^s_{FF}+d^s_{E|F} \quad \text{where} \quad d^s_{E|F}=\frac{1}{m}\left[\sum_C\sum_{K_{\bar{C}}}\text{Loss}(\langle E_C K_{\bar{C}}\rangle,\langle F_C K_{\bar{C}}\rangle)-\sum_{u=1}\text{Loss}(x_{Eu},x_{Fu})\right]$$

and where the sum over C refers to all subsets of variables in C whose values differ between E and F.

Step 3: Search for $\min_{E,F}(d^s_{EF})$ in D^s and print or store on a history record: s+1, D^s if desired, $\min_{E,F}(d^s_{EF})$, and for this minimum: E, F, the newly assigned label G, and for all values $x_{Eu}\neq x_{Fu}$: x_{Eu}, x_{Fu}, x_{Gu}, u.

Step 4: Merge E and F into G by modifying for all $x_{Eu}\neq x_{Fu}$ and u:

$$n_{x_{Gu}}=n_{x_{Eu}}+n_{x_{Fu}}, \qquad n_{x_{Eu}}=n_{x_{Fu}}=0$$

for all clusters whose objects share some value x with values in $K_{\bar{C}}$ of E and F:

$$n_{\langle G_C K_{\bar{C}}\rangle}=n_{\langle E_C K_{\bar{C}}\rangle}+n_{\langle F_C K_{\bar{C}}\rangle}, \qquad n_{\langle E_C K_{\bar{C}}\rangle}=n_{\langle F_C K_{\bar{C}}\rangle}=0$$

and

$$d_{GG}=d_{EF}, \qquad d_{EI}=d_{IF}=0 \text{ for all } I\neq E,F$$

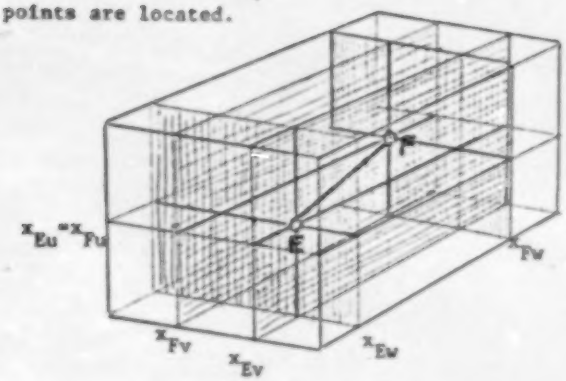
recompute s so that the number of unique objects or clusters in n-s.

Step 5: Return altered accounts to step 2 unless either $\max(d_{GG})$ exceeds a specified limit, the number of clusters n-s falls below a specified number, or any other terminating criterion is satisfied.

Step 6: Print results and terminate.

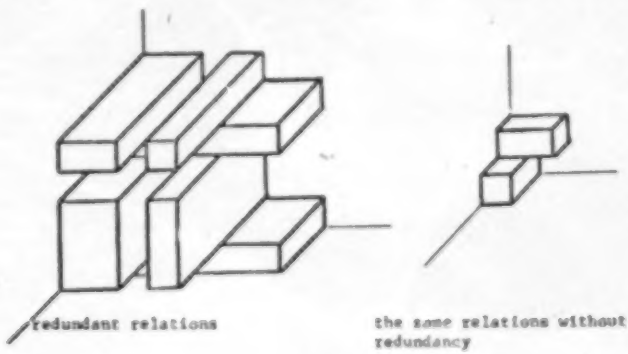
Note that d_{EF} measures the amount of structure lost by merging E and F. Within the three dimensions u, v, and w, if $x_{Eu}=x_{Fu}$ are the values E and F share, the distance between E and F expresses the difference it would make to the total amount of structure in the data when (with K denoting clusters other than E or F) all triples $\langle x_{Ku},x_{Kv},x_{Ew}\rangle$ and $\langle x_{Ku},x_{Ku},x_{Fu}\rangle$, $\langle x_{Ku},x_{Ev},x_{Kw}\rangle$ and $\langle x_{Ku},x_{Fv},x_{Kw}\rangle$, and $\langle x_{Ku},x_{Ev},x_{Ew}\rangle$ and $\langle x_{Ku},x_{Fv},x_{Fw}\rangle$ would no longer be differentiated. The sum over C then assures that all clusters are merged whose objects share some value with E or F in v, in w and in both vw. What d_{EF} assesses is the effect of collapsing not only the point E and F but also all planes on

which these points are located.



The implementation of the previously published version of the multivariate classification algorithm used an extremely wasteful form of storage (objects occupied cells in an m-dimensional array) and more complicated accounting devices that made the procedure reach computational limits before practical results could be obtained. The above algorithm is currently under investigation.

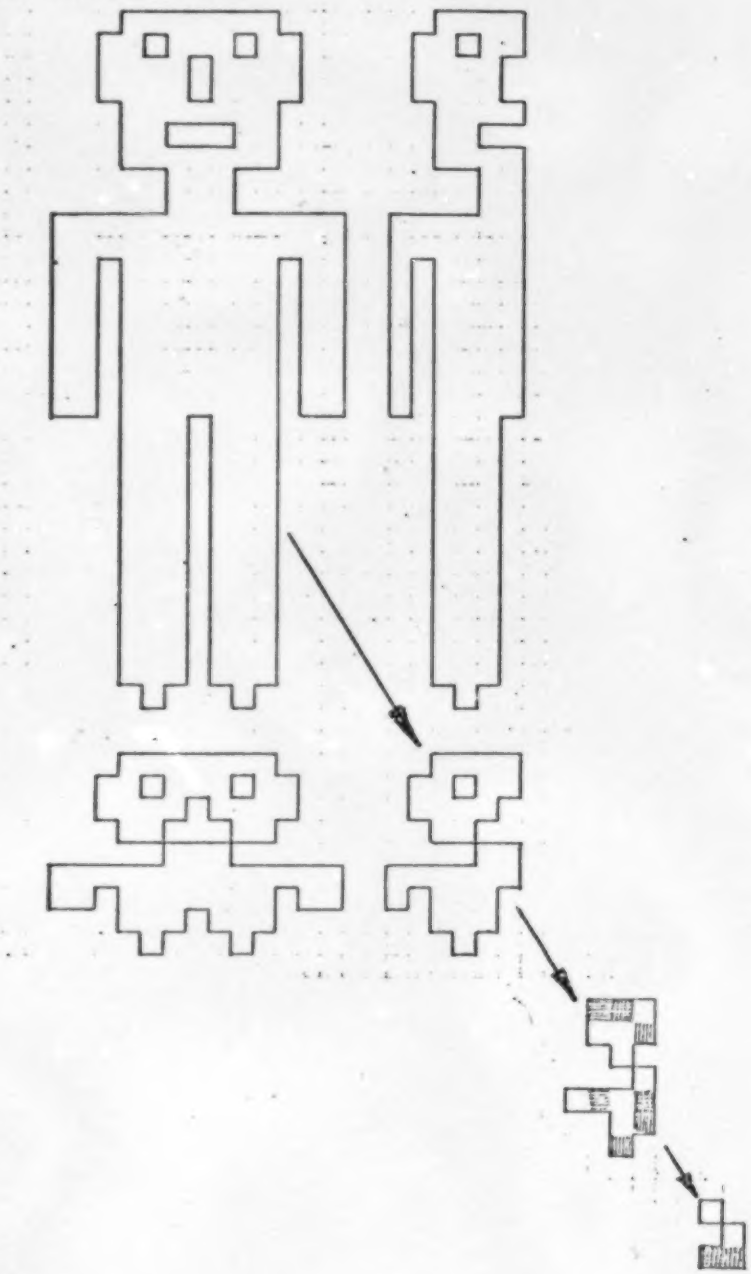
What either algorithm for multivariate classification accomplishes might best be illustrated graphically. Suppose a three-valued characterization of a sample of objects finds all objects distributed as in the left space. There is a lot of redundancy in the values used for describing these objects and there is also some structure manifest in the distribution. Multivariate classification would now attempt to eliminate this redundancy by grouping variables in such a way that the remaining space contains as much of the initial structure as possible. In the illustration on the right of the original distribution no structure is lost. The algorithm boiled the initial representation down to its essentials.



redundant relations

the same relations without redundancy

In another, somewhat more artificial example, suppose the schematic figure of a man is given within two dimensions:



The clustering of values in the horizontal dimension first eliminates the duplication of columns, here due to symmetry, and yields the figure to the right of the original, showing no loss. The clustering of values in vertical dimension eliminates all duplication of rows and yields the figure below the original, showing no loss either. Clustering in both dimensions yields the resultant figure below and right of the original, also showing no loss in structure. (At this point it might be said that the example is insofar misleading as the algorithm does not recognise proximities between rows and columns which are important to recognise the gestalt. But, except for this fact,) the original figure can be reconstructed from the figure below and right of the original by inverse application of the hierarchical clustering that emerged in each variable.

Further simplification of the figure results in losses which are indicated here by shading. When such losses occur they cannot necessarily be evaluated from either dimension in isolation. Losses in structure are losses in interaction effects and thus require that losses be inspected simultaneously for all variables involved. Continuing the, here bivariate, classification one might reach the figure in the extreme lower right corner. Any two more steps will then wipe out the structure that is still manifest in this simple figure. Where to stop the clustering is a question of applying suitable termination criteria on the process.

It should be mentioned that a variety of clustering algorithms are published in Hartigan (1975) whose work was published after this was substantially completed.

Validation and Vindication

Clustering procedures compute clusters, often regardless of how strong the patterns permeate the data these clusters aim to represent. In any distribution of objects in space, even entirely random ones, some objects are bound to be closer to or more similar than others. Even in rare uniform distributions, the slightest inequality, can provide the kick that starts a clustering sequence rolling. Testing for the statistical significance of the clusters that do emerge in the process therefore is an important safeguard agsinst attempts to interpret the results of a clustering process when an underlying pattern is spurious or does not exist in fact. The minimum requirement is to test one of two null-hypotheses: that the co-occurring qualities shared within a cluster are due to chance or that the objects of different clusters are drawn from the same population.

But evidence for the statistical significance of clusters should be seen only as a prerequisite for entering into validity considerations. Only when null-hypotheses can be rejected with confidence might one find empirically meaningful interpretations of clustering results. When clustering turns out to be due to chance or an artefact of the procedure, their potential validity is out of the question.

Although I am quite aware of existing typologies for kinds of validation, let me focus here on only two types that Feigl (1952) termed validation and vindication. In the context of this application, validation is a mode of justification according to which the results of a particular analytical procedure are justified by showing the structure of that procedure to be derivable from general principles or theories that are accepted quite independently of the procedure to be validated; while vindication is a mode of justification that renders a particular analytical procedure acceptable on the grounds that its results lead to accurate predictions (to a degree better than chance) regardless of the details of the procedure. The rules of deduction and induction are essential to validation while the relation between means and particular ends provide the basis for vindication. In focussing on these two kinds of justification, I take for granted that the procedure is reliable, that successive clustering is order-invariant, that distances

and homogeneity measures satisfy required conditions, etc., all of which can be justified on logical grounds. I also take for granted that data are relevant in the sense considered earlier for it is inconceivable that valid clusters can be obtained from irrelevant data.

Two not necessarily separate questions pertain to the validation of clustering procedures. First, exactly what features of objects are characterised when data enter the procedure in their derived form as distance or similarity measures and whether these features and the omission of others can be justified on theoretical or on empirical grounds. And, second, exactly what does a clustering procedure optimize, the clustering criteria it employs and how the measures that characterise the emerging clusters relate to a theory about how objects become associated, group themselves or are clustered in reality.

Regarding the first question, it should be noted that there are great differences between how product-moment correlations, Euclidian distances or information losses conceptualize and quantitatively assess dissimilarities between objects. For example, the product-moment correlation assesses the degree to which two objects are linearly related. A positive r_{ij} indicates that the values of two objects increase in the same direction while a negative r_{ij} indicates such an increase to be in the opposite direction. But $r_{ij} = +1$ does not imply $i=j$. The underlying concept of resemblance is a very peculiar one and the researcher who wishes to cluster on the basis of a correlation matrix must establish that available knowledge about the nature of the objects would indeed lead to this conception. Product-moment correlations assume that objects i and j are related by $a_j x_{ju} + b_j = a_i x_{iu} + b_i$ for all variables u and that similarity is independent of the constants a and b. Obviously this does not conform to intuition which suggests that two objects are maximally similar only when $i=j$, that is, $x_{ju} = x_{iu}$ for all of the u, a condition that Pearson's intra-class correlation coefficient satisfies, but not the product-moment coefficient.

An eximination of whether the formal properties of the underlying distance or similarity measures are defensible ought to be made before any clustering for a particular purpose is undertaken. Failure to provide such validating evidence make it otherwise difficult to interpret findings in the light of a given theory.

One common way of by-passing the validation of distances is to input data in the form of subjective difference or similarity ratings obtained from a sample of subjects. If the researcher is indeed interested in clustering subjective difference or similarity judgements, two problems tend to arise. One i_ the variance that such judgements invariably entail and the other is that such judgements too ought to satisfy the formal conditions of a distance metric.

Answers to the second question pertaining to the validation of clustering criteria are an even more important problem for distances are merely the most reduced cases of difference or heterogeneity measures. To decide on the acceptability of a clustering criteria, the researcher must first of all decide on or know what the clusters are to represent. Given such a designation of purpose the researcher must then examine the principles underlying the formation of the groupings in reality that a clustering attempts to approximate or predict. The consistency between the decision criteria, difference or heterogeneity measures employed in the clustering procedure on the one side and knowledge about the natural processes on the other must then be demonstrated to complete the validation. This knowledge may take the form of an established theory, of hard empirical evidence or, in its weakest form, of grounded intuition. Wherever such knowledge comes from, validation relies on it.

For example, if the resulting clusters are expected to predict how individuals form cliques or other social forms of organization, then knowledge about the way such social groupings emerge is indispensable in the validation of a procedure. The knowledge that cliques and social groups possess synergetic-organizational-gestalt qualities and that their formation cannot be predicted from information on the interaction within pairs of individuals would render bi-ordinal techniques invalid from the start (unless the effect

is insignificantly small). To be valid, a multi-ordinal technique would then have to replicate the social process involved.

Another crucial option is whether clusters are formed on the basis of differences between clusters or heterogeneities within. The chain-like clusters resulting from the connectedness method has already been contrasted with the compactness of the diameter method. The knowledge that all similarities or distances within a cluster will determine its boundary relative to other clusters lends validity to a minimum heterogeneity criterion, whereas the knowledge that significant similarities and distances follow a hierarchical pattern, representing differences between clusters while neglecting those within would lend validity to a minimum difference criterion. A hypothetical example of a situation in which a difference measure might be superior to a heterogeneity measure would be a certain form of communication within a social organization in which communication occurs primarily on the same level of the organizational hierarchy and between minimally different parts of the organization and is secondary or absent across different levels of such a hierarchy and within these parts. Such an implicit hierarchical conception of difference would be inappropriate when an organization is formed on the basis that members share certain properties, that communication within is larger or more important that communication across the parts of an organization, etc.

It is often more difficult to apply available evidence about natural groupings on available clustering techniques then to formalize such evidence into a computable clustering criterion. The development of the strong association technique by Krippendorff (1975) is a case in point. It started with a problem in content analysis where the development of emic categories is a common problem. The task was to develop a reliable coding instrument for advertising appeals in categories that are close to those used by television viewers. For this purpose Dziurzynski (1976) asked subjects to group about 300 appeals culled from commercials into categories that seemed most meaningful to them. In observing the subjects' justifications one often finds some like this: "if i and j are together then k must be in the same category but if h and i are in the same category then k cannot join them." Those are typical multi-ordinal arguments. The task was to form clusters

among aspects based on agreements among subjects regarding the grouping. The formalization of the motion of agreement which ought to be maximum when groups are either identical or when one is included in the other, lead to a generalization of Benini's association coefficient which has been discussed above. The correspondence was taken as validating evidence.

To summarize, validation asks whether the way information is processed within a clustering procedure is consistent with the way such information would be processed in the real world while vindication asks the a posteriori question of whether the results of a clustering procedure correspond with independently obtained evidence about clusters.

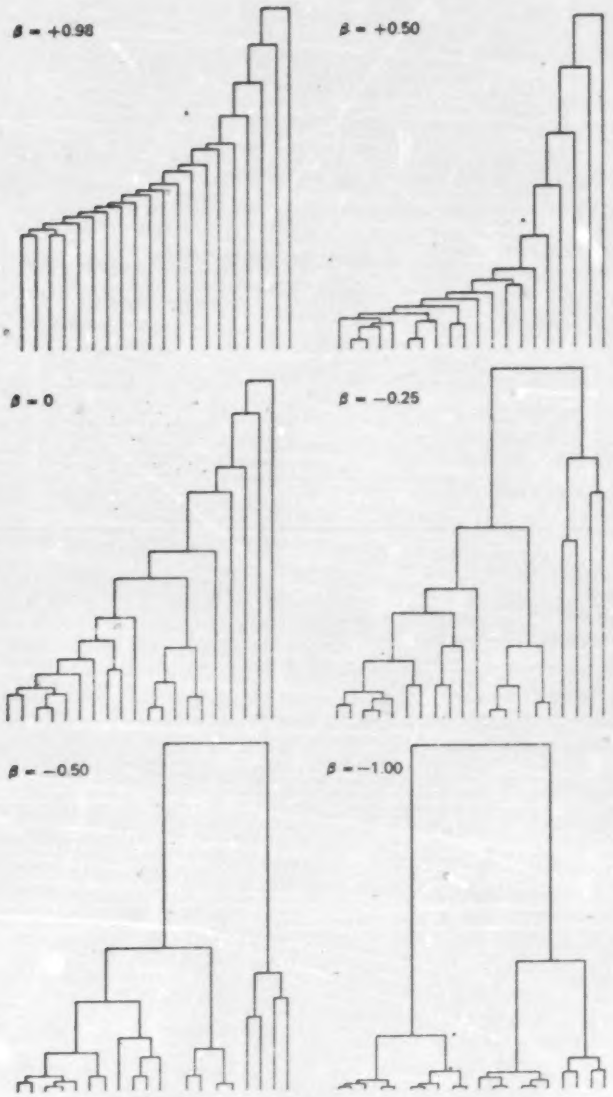
The most obvious form of vindication is to establish correspondence between the results of a clustering procedure and the results obtained by other methods (including by independent observation). Since clusters obtained by other methods must always be available for such comparisons, vindication primarily yields information about the efficiency or simplicity, and only secondarily about the adequacy of the underlying structure.

So, when developing the partitioning algorithm CONCOR and probably because correlations of correlations of ... is a concept that is far removed from penetrations by intuition, Breiger, Boorman and Arabie (1975) compared their results with those obtained by a variety of other clustering techniques. The finding that CONCOR results approximate those obtained by the single linkage method makes the procedure vindicatively acceptable but only to the extent the results of this single linkage method are already known to be valid in a particular application.

In another example of vindication, we asked subjects to group sets of words according to perceived semantic similarities. Since the multivariate classification algorithm was developed by formalizing certain theories of contextual meanings, if the theories and their algorithmization are correct, then computational results and subjective clusters are expected to be in high agreement. In this case we were fortunate to be able to vary certain computational parameters and found, to our surprise, that the weakest clustering

criterion resulted in the best fits. This may serve as a warning against the assumption that the validity of clustering procedures increases with their complexity.

In vindication experiments, the variability of a clustering technique is a deceptive virtue, however, for it is always possible to find a computational approximation to an independently obtained set of clusters. This possibility is exemplified in work done by Lance and Williams (1967) who showed how changes in value of one variable of his clustering



criterion causes extremely different dendrograms to emerge from the same data. The danger is that, once one considers himself free to play with the clustering criteria, one can "prove" anything, and, since the computation then merely supports what is already known, the proof given is "empty."

Vindication allows all conceivable clustering options to be tested against empirical evidence but its aim is to find that option which produces consistently high agreements within the empirical domain chosen. A single "convincing match" means very little. Carefully used, vindication provides a method for generalising or for confining the success of a particular clustering procedure.

The researcher who does not have independently obtained clusters at his disposal might be lead to believe that the computational results "make sense" or are "acceptable on intuitive grounds." But a better way of rendering such results plausible is to get into the very procedure that produced them and to show that the procedure is a homomorphic representation of the processes known to explain the phenomena under consideration. All users of clustering techniques should be expected to make at least some effort at validation when publishing their results.

References

Bailey, K.D. Cluster Analysis. Pp. 59-128 in Heise, D.R. (Ed.) Sociological Methodology 1975. San Francisco : Jossey-Bass, 1974.

Benini, R. Principii di Demografia. No.29. of Manuali Barbèra de Scienze Guiviche Sociale e Politiche, Firenzi: G. Barbera, 1901.

Bock, H.H. Automatische Klassifikation, Göttingen: Vandenhoeck & Ruprecht, 1974.

Breiger, R.L., Boorman, S.A. and Arabie, P. An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimentional Scaling. Journal of Mathematical Psychology, 12 : 328 - 383, 1975.

Cureton, H.R., Cureton, L.W. and Durell, R.C. A method of Cluster Analysis. Multivariate Behavioral Research, 5: 101-106, 1970.

Driver, H.E. and Kroeber, A.L. Quantitative Expression of Cultural Relationships. University of California Publications in American Archeology and Ethnology, 31: 211-256, 1932.

Dziurzynski, P. Masters thesis in progress, Philadelphia: The Annenberg School of Communication, University of Pennsylvania.

Feigl, H. Validation and Vindication: An Analysis of the Nature and Limits of Ethical Arguments. Pp. 667-680 in Sellars W. and Hospers J. (Eds.) Readings in Ethical Theory. New York: Appleton-Century - Crofts, Inc., 1952.

Gower, J.C. and Ross, G.J.S. Minimum Spanning Trees and Single-Linkage Cluster Analysis. Applied Statistics, 18: 54-64, 1969.

Gower, J.C. A General Coefficient of Similarity and some of its Properties. Biometrics, 27: 857-872, 1971.

Hamming, R.W. Error Detecting and Error Correcting Codes. The Bell System Technical Journal, 26,2: 147-160, 1950.

Hartigan, John A. Clustering Algorithms., New York: Wiley, 1975.

Heinecke, F. Naturgeschichte des Herings. I. Die Lokalformen und die Wanderungen des Herings in den Europäischen Meeren. Abhandlungen des Deutschen Seefische ei-Vereins 2: i-cxxxvi, 1-223, 1898.

Jardine, N. and Simbson R. Mathematical Taxonomy. New York: Wiley, 1971.

Johnson, S.C. Hierarchical Clustering Schemes. Psychometrika, 32: 241-254, 1967.

King, B. Step-wise Clustering Procedures. Journal of the Am. Stat. Ass., 62: 86-101, 1967.

Krippendorff, K. Computer Programs for Multivariate Classification in Content Analysis, A Proposal to the National Science Foundation, Philadelphia, University of Pennsylvania, mimeo 1969.

Krippendorff, K. Reliability. Philadelphia: University of Pennsylvania, mimeo, 1973.

Krippendorff, K. An Algorithm of Simplifying the Representation of Complex Systems, pp. 1693-1702 in John Rose (Ed.), Advances in Cybernetics and Systems, New York: Gorden & Breach, 1974.

Krippendorff, K. A method for Strong Associative Clustering of 2^m Data. Philadelphia: University of Pennsylvania Mimeo, 1975.

Lance, G.N. and Williams, W.T. A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems. Computer J., 9: 373-380, 1967.

Lance, G.N. and Williams, W.T. A General Theory of Classificatory Sorting Strategies II. Clustering Systems. Computer J., 10: 271-277, 1967.

Lorr, M.A. A Review and Classification of Typological Procedures. Paper Read at the meeting of the American Psychological Association, San Francisco, 1968. Cited in Careton, Careton and Durell, 1970.

Lysenko, O., and Sneath, P.H.A. The Use of Models in Bacterial Classification. Journal of General Microbiology, 20: 284-290, 1959.

Macnaughton-Smith, P. Some Statistical and Other Numerical Techniques for Classifying Individuals. Home Office Research Unit Report No. 6, London: H.M. S.O., 1965.

Mahalanobis, P.C. On the Generalized Distance in Statistics. Proc. Nat. Inst. Sci. India, 2: 49-55, 1936.

Miller, G.A. and Nicely, P.E. An Analysis of Perceptual Confusion Among Some English Consonants. J. Acoust. Sol. America, 27: 338-352, 1955.

Osgood, C.E., Suci, J.C. and Tannenbaum, P.H. The Measurement of Meaning, Urbana: University of Illinois Press, 1957.

Pearson, K. On the Coefficient of Radical Likeness. Biometrica, 18: 105-117, 1926.

Rao, C.R. The Utilization of Multiple Measurements in Problems of Biological Classification. Journal of the Royal Statistical Society, B10: 159-193, 1948.

Rao C.R. Advanced Statistical Methods in Biometric Research, New York: Wiley, 1952.

Sneath, P.H.A., and Sokal, R.R. Numerical Taxonomy, San Francisco: Freeman, 1973.

Socal, R.R. and Sneath, P.H.A. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman, 1963.

Tyron, R.C. Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Ann Arbor: Edwards Brothers, 1939.

Tyron, R.C. and Bailey, D.E. Cluster Analysis New York: McGraw-Hill, 1970.

Zubin, J.A. A Technique for Measuring Likemindedness. J. of Abn. Soc. Psych., 33: 508-516, 1938.